INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

Certificate course on AI and ML

A Refresher on Mathematics

Prof. Kusum Deep
Full Professor (HAG), Department of Mathematics
Joint Faculty, Centre for Artificial Intelligence & Data Science
Indian Institute of Technology Roorkee, Roorkee – 247667

kusum.deep@ma.iitr.ac.in, kusumdeep@gmail.com

Learning outcomes

- Overview of Linear Algebra
- Matrices and their basic operations
- Rank, Eigen values and Eigen vectors
- Singular Value Decomposition
- Principal Component Analysis
- Concept of gradient
- Partial derivatives and Chain rule
- Optimization

Algebra is the set of objects (or symbols) and a set of rules to manipulate these objects.

Linear algebra is the study of vectors and certain algebraic rules to manipulate vectors.

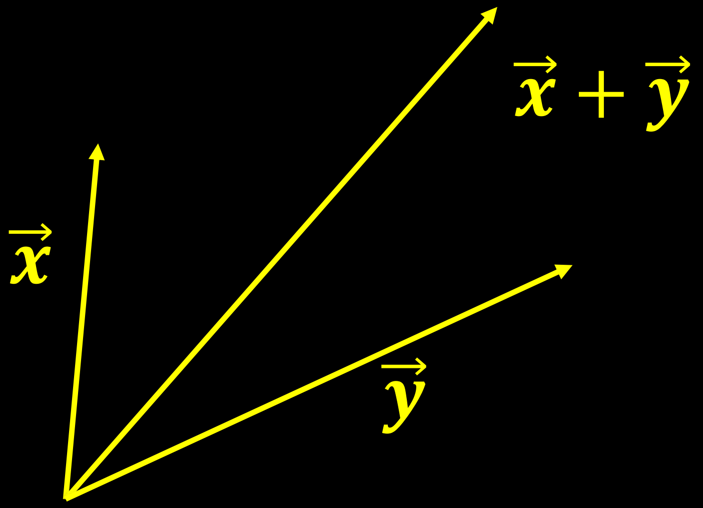
Vectors are special objects that can be added together and multiplied by scalars to produce another object of the same kind.

Examples: **Geometric vectors**

Let $\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}$ be two vectors (in R^2).

Then, $\vec{\boldsymbol{z}} = \vec{\boldsymbol{x}} + \vec{\boldsymbol{y}}$ is also a vector.

Also, $\lambda\vec{\boldsymbol{x}}$ is also a vector, where $\lambda \in R$.



Ex 2: Polynomials over real numbers:

$$P(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

$$\text{Q}(x) = b_0 + b_1 x + b_2 x^2 + \cdots + b_m x^m$$

Then $\mathrm{R(x)} = P(x) + Q(x)$

$= a_0 + b_0 + (a_1 + b_1)x + (a_2 + b_2)x^2 + \cdots + (a_n + b_n)x^n + \cdots + b_m x^m$ assuming $m > n$

And $\lambda P(x) = \lambda a_0 + \lambda a_1 x + \lambda a_2 x^2 +..+ \lambda a_n x^n$ where λ is a real number.

Ex 3: Audio signals

Most Commonly used vectors are in R^n , n is a positive integer. (R is real numbers) e.g. R^2.

$(2, 8), (-5, 3), \dots .$ Represent vectors in R^2.

In R^n, any tuple of the type (x_1, $x_{2,}$ … x_n) is vector.

Magnitude of a vector X = (x_1, $x_{2,}$ … x_n) is

$$\|X\| = \sqrt{x_1^2 + x_2^3 + \ \dots \ x_n^2}$$

Zero vector all components are zero.

Standard vectors are R^n are:

(1, 0, …0), (0,1, … 0), (0,0,… 1).

Algebra of vectors

Let $X = (x_1, x_2, \ldots\ x_n)$ and $Y = (y_1, y_2, \ldots\ y_n)$ be any vectors in R^n, then:

1. $|X| = \sqrt{x_1^2 + x_2^3 + \ \ldots\ x_n^2}$
2. $X = Y$ iff $x_1 = y_1, x_2 = y_2, \ldots\ x_n = y_n$
3. $X \pm Y = (x_1 \pm y_1, x_2 \pm y_2, \ldots\ x_n \pm y_n)$
4. $X.Y = x_1.y_1 + x_2.y_2 + \cdots\ x_n.y_n$
5. Dot product $X.Y = |X||Y| \cos\theta$, where θ is angle between X and Y.
6. Cross product $X \times Y = |X||Y| \sin\theta$, where θ is angle between X and Y.

System of equations: Ex 1

$$x_1 + x_2 + x_3 = 3 \ldots \ldots \ldots \ldots \ldots \ldots \ldots (1)$$
$$x_1 - x_2 + 2x_3 = 2 \ldots \ldots \ldots \ldots \ldots \ldots \ldots (2)$$
$$2x_1 \qquad + 3x_3 = 1 \ldots \ldots \ldots \ldots \ldots \ldots \ldots (3)$$

Adding (1) and (2) , gives:

$$2x_1 \qquad + 3x_3 = 5$$

This contradicts (3)

So, this system of equations has **<u>no solution</u>**

Ex 2

$$x_1 + x_2 + x_3 = 3 \ldots \ldots \ldots \ldots \ldots \ldots \ldots (1)$$
$$x_1 - x_2 + 2x_3 = 2 \ldots \ldots \ldots \ldots \ldots \ldots \ldots (2)$$
$$x_2 + x_3 = 2 \ldots \ldots \ldots \ldots \ldots \ldots \ldots (3)$$

Substituting (3) in (1), gives $x_1 = 1$

Adding (1) and (2), gives:

$2x_1 + 3x_3 = 5$, i.e. $x_3 = 1$.

This system of equations has **unique solution.**

(1, 1, 1)

Ex 3

$$x_1 + x_2 + x_3 = 3 \dots \dots \dots \dots \dots \dots \dots (1)$$
$$x_1 - x_2 + 2x_3 = 2 \dots \dots \dots \dots \dots \dots \dots (2)$$
$$2x_2 + 3x_3 = 5 \dots \dots \dots \dots \dots \dots \dots (3)$$

Adding (1) and (2), gives (3). So we can omit (3).

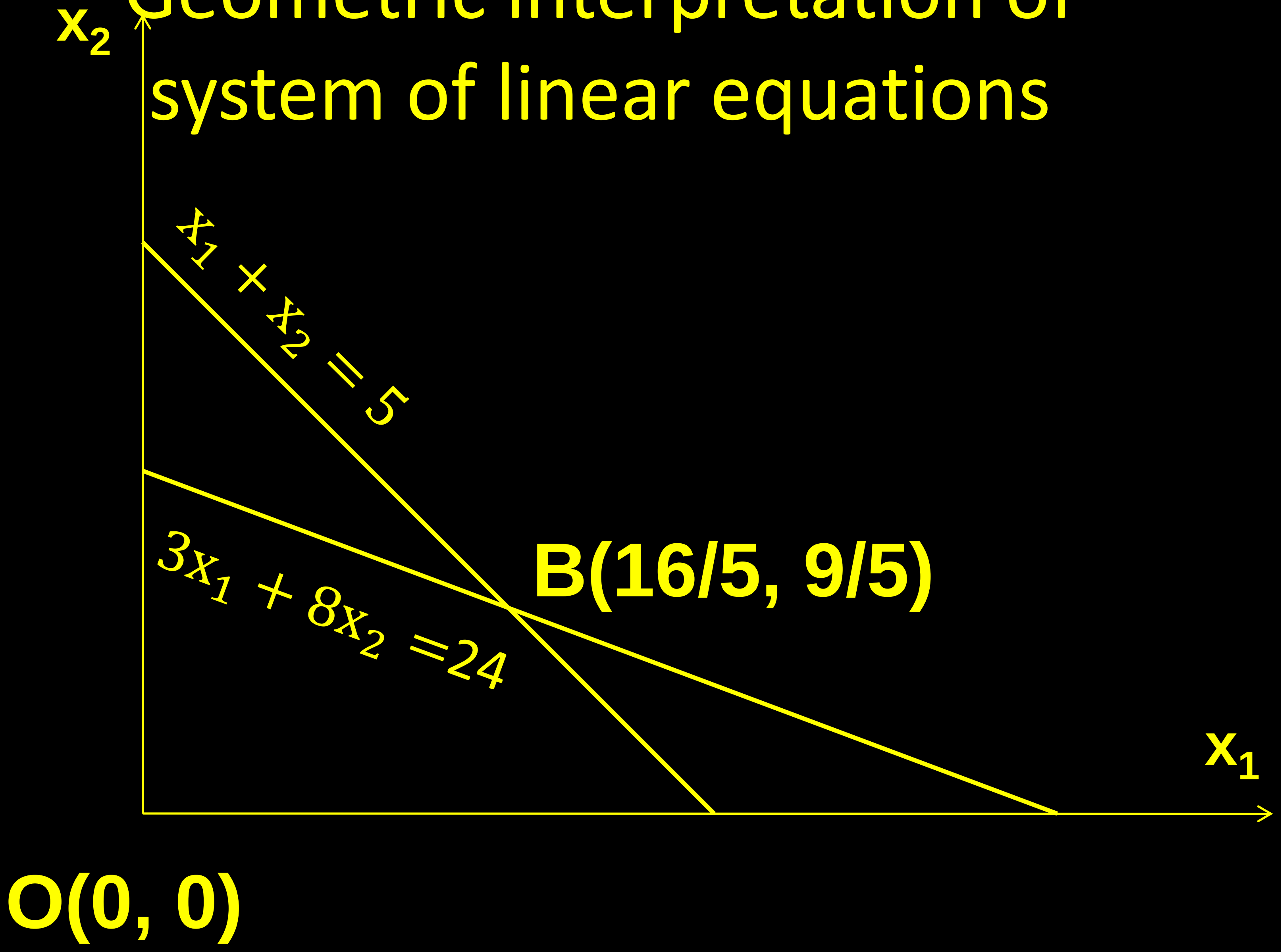
From (1) and (2), we get:

$2x_1 = 5 - 3x_3$ and $2x_2 = 1 + x_3$

This system has **infinite no. of solution.**

$$\left(\frac{5}{2} - \frac{3}{2}a, \frac{1}{2} + \frac{1}{2}a, a\right) for\ any\ a \in R$$

Geometric interpretation of system of linear equations



System of linear equations

$$\boldsymbol{Ax} = \boldsymbol{b}$$

$$\Leftrightarrow \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} x_1 + \begin{bmatrix} a_{12} \\ \vdots \\ a_{m2} \end{bmatrix} x_2 + \ldots \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix} x_n = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

Matrices

Def: With m and n as natural numbers, a real-valued (m, n) *matrix* $\boldsymbol{A}$ is a m-n-tuple of elements a$_{ij}$, i = 1, 2, … m, j = 1, 2, … n, which is ordered according to a rectangular scheme consisting of m rows and n columns.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$, where a_{ij} are reals.

Ex: $\boldsymbol{P} = \begin{bmatrix} 2 & 5 & 1 \\ -1 & 1.5 & 0.82 \end{bmatrix}$

Rael life examples:

Students sitting in a class:

1st row: Tim, Ravi, Rohit

2nd row: Amar, Ram

3rd row: Amit, Irfan, John

This information can be expressed in matrix:

$$A = \begin{bmatrix} Tim & Ravi & Rohit \\ Amar & Ram & \\ Amit & Irfan & John \end{bmatrix}$$

Equal matrix

Let $\boldsymbol{A} = \left(a_{ij}\right)$ and $\boldsymbol{B} = \left(b_{ij}\right)$ be two matrix (over real numbers) of size $m \times n$.

Then $\boldsymbol{A} = \boldsymbol{B}$ if $a_{ij} = b_{ij}$

for all $i = 1, 2, \dots m$; $j = 1, 2, \dots n$.

Ex: $\boldsymbol{A} = \begin{bmatrix} 2 & 4 \\ 1 & 7 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} 2 & 4 \\ 1 & 7 \end{bmatrix}$

Identity matrix of order n is: $\boldsymbol{I} = \begin{bmatrix} 1 & \dots & 0 \\ \vdots & 1 & \vdots \\ 0 & \dots & 1 \end{bmatrix}$

Ex: $\boldsymbol{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ is an identity matrix of order 3

Null matrix is having all entries as **0**.

Ex: $\mathbf{P} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ is a null matrix of order 3.

Matrix is a **Diagonal matrix** if $A = \left(a_{ij}\right) = 0 \quad if\ i \neq j$

Ex: $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & -4 \end{bmatrix}$, is a diagonal matrix of order 3.

Matrix is **upper triangular** if $A = \left(a_{ij}\right) = 0 \quad if\ i > j$

Ex: $\begin{bmatrix} 1 & 7 & 4.5 \\ 0 & 5.5 & 0 \\ 0 & 0 & -4 \end{bmatrix}$, is a upper triangular matrix of order 3.

Matrix is **lower triangular** if $A = \left(a_{ij}\right) = 0 \quad if\ i < j$

Ex: $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 8 & 5.5 & 0 & 0 \\ 3 & 3 & 4 & 0 \\ 1 & 8 & 2 & 5 \end{bmatrix}$, is a lower triangular matrix of order 4.

Addition of matrices

If A and B are mXn matrices, then their sum is also a mXn matrix

Let $\boldsymbol{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} b_{11} & \dots & b_{1n} \\ \vdots & & \vdots \\ b_{m1} & \dots & b_{mn} \end{bmatrix}$

Then, $\boldsymbol{A} + \boldsymbol{B} = \begin{bmatrix} a_{11} + b_{11} & \dots & a_{1n} + b_{1n} \\ \vdots & & \vdots \\ a_{m1} + b_{m1} & \dots & a_{mn} + b_{mn} \end{bmatrix}$

Ex: If $\boldsymbol{A} = \begin{bmatrix} 2 & 0 & 1 \\ -1 & 1.5 & 0.8 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 1 & 5 & 0 \\ 3 & -1.5 & 0.7 \end{bmatrix}$,

Then $\boldsymbol{A} + \boldsymbol{B} = \begin{bmatrix} 2 + 1 = 3 & 5 & 1 \\ 2 & 0 & 1.5 \end{bmatrix}$

Multiplication of matrices

If $\boldsymbol{A}$ is mXn matrix and $\boldsymbol{B}$ is a nXp matrix, then AB is a mXp matrix.

Let $\boldsymbol{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} b_{11} & \dots & b_{1p} \\ \vdots & & \vdots \\ b_{n1} & \dots & b_{np} \end{bmatrix}$

$$\boldsymbol{AB} = \begin{bmatrix} a_{11}b_{11} + \cdots + a_{1n}b_{n1} & \dots & a_{11}b_{1p} + \cdots + a_{1n}b_{np} \\ \vdots & & \vdots \\ a_{m1}b_{11} + \cdots + a_{mn}b_{n1} & \dots & a_{m1}b_{1p} + \cdots + a_{mn}b_{np} \end{bmatrix}$$

Ex: If $\boldsymbol{A} = \begin{bmatrix} 2 & 0 & 1 \\ -1 & 1.5 & 0.8 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} 1 & 2 \\ -2 & -1 \\ 3 & 4 \end{bmatrix}$,

Then $\boldsymbol{AB} = \begin{bmatrix} 2X1 + 0X(-2) + 1X3 = 5 & 8 \\ -1.6 & -0.3 \end{bmatrix}$ and

$$\boldsymbol{BA} = \begin{bmatrix} 4 & 3 & 2.6 \\ -5 & -1.5 & -2.8 \\ 2 & 6 & 6.2 \end{bmatrix}$$

Ex: : If $\boldsymbol{A} = \begin{bmatrix} 2 & 9 & -3 \\ 5 & -1 & 6 \\ 7 & 2 & 4 \end{bmatrix}$ and

$$\boldsymbol{B} = \begin{bmatrix} 9 & -5 & 2 \\ 7 & 4 & 1 \\ 8 & 3 & 6 \end{bmatrix}$$

Find $A \times B$

Also find $B \times A$

Remarks:

1. $\boldsymbol{AB \neq BA}$
2. If $\boldsymbol{AB} =$**0**, it is not necessary that **A** or **B** is **0**.

   Ex: $\boldsymbol{A} = \begin{bmatrix} 1 & -1 \\ 3 & -3 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}$
3. If $\boldsymbol{AB} = \boldsymbol{AC}$, then it is not necessary that $\boldsymbol{B} = \boldsymbol{C}$

   Ex: $\boldsymbol{A} = \begin{bmatrix} 2 & -3 \\ -4 & 6 \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} 8 & 4 \\ 5 & 5 \end{bmatrix}$ and $\boldsymbol{C} = \begin{bmatrix} 5 & -2 \\ 3 & 1 \end{bmatrix}$

   Then $\boldsymbol{AB} = \begin{bmatrix} 1 & -7 \\ -2 & 14 \end{bmatrix}$ and $\boldsymbol{AC} = \begin{bmatrix} 1 & -7 \\ -2 & 14 \end{bmatrix}$

Associativity

For all $\mathbf{A}_{mXn}$, $\mathbf{B}_{nXp}$ and $\mathbf{C}_{pXq}$ then $\mathbf{(AB)C=A(BC)}$

Distributivity

For all $\mathbf{A}_{mXn}$, $\mathbf{B}_{mXn}$ and $\mathbf{C}_{nXp}$, $\mathbf{D}_{nXp}$, then:

$$\mathbf{(A + B)C = AC + BC}$$

$$\mathbf{A(C +D) = AC + AD}$$

Multiplication with identity

For all $\mathbf{A}_{mXm}$, $\mathbf{I}_m\mathbf{A} = \mathbf{AI}_m\mathbf{= A}$

Inverse of a matrix

Def: Let $\boldsymbol{A}$ be a square matrix of order n. Then a square matrix $\boldsymbol{B}$ of order n is said to be the **inverse of $\boldsymbol{A}$** if

$$\boldsymbol{AB = I_n = BA}$$

It is denote by $\boldsymbol{A^{-1}}$.

Notes:

1. Not all matrices are invertible.
2. Inverse of a matrix is unique.

Existence of inverse of 2X2 matrix

$Let\ \boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, and $\mathbf{B} = \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$

$$\boldsymbol{AB} = \begin{bmatrix} a_{11}a_{22} - a_{12}a_{21} & 0 \\ 0 & a_{11}a_{22} - a_{12}a_{21} \end{bmatrix}$$
$$= (a_{11}a_{22} - a_{12}a_{21})I_n$$

Therefore, $\boldsymbol{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$

If and only if $a_{11}a_{22} - a_{12}a_{21} \neq 0$

That is **determinant** is non zero. These are called **non-singular matrix.**

Example

Let $\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 1 \\ 4 & 4 & 5 \\ 6 & 7 & 7 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} -7 & -7 & 6 \\ 2 & 1 & -1 \\ 4 & 5 & -4 \end{bmatrix}$

Are inverse of each other since $\boldsymbol{AB} = \boldsymbol{I} = \boldsymbol{BA}$.

Transpose of a matrix

Def: If $\boldsymbol{A}$ is a mXn matrix, then the nXm matrix $\boldsymbol{B}$ is said to be **transpose** of $\boldsymbol{A}$, if $b_{ji} = a_{ij}$.

It is denoted by $\boldsymbol{A}^T$.

Ex:

let $\boldsymbol{A} = \begin{bmatrix} 2 & 5 & 1 \\ -1 & 1.5 & 0.82 \end{bmatrix}$,then $\boldsymbol{A}^T = \begin{bmatrix} 2 & -1 \\ 5 & 1.5 \\ 1 & 0.82 \end{bmatrix}$

Properties of inverse and transpose

$$\boldsymbol{AA^{-1} = A^{-1}A = I}$$

$$\boldsymbol{(AB)^{-1} = B^{-1}A^{-1}}$$

$$\boldsymbol{(A+B)^{-1} \neq A^{-1} + B^{-1}}$$

$$\boldsymbol{(A^T)^T = A}$$

$$\boldsymbol{(A+B)^T = A^T + B^T}$$

$$\boldsymbol{(AB)^T = B^T A^T}$$

$$(k\boldsymbol{A})^T = k\boldsymbol{A}^T$$

Symmetric Matrix

Def: A square matrix of order n is said to be **symmetric** if $A = A^T$.

Ex: $\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 6 \\ 3 & 6 & 9 \end{bmatrix}$

Notes:

1. If **A** is invertible then $\boldsymbol{A}^T$ is also invertible.
2. $(\boldsymbol{A}^{-1})^T = (\boldsymbol{A}^T)^{-1}$

The sum of symmetric matrices is always symmetric.

But the product of symmetric matrices need not be symmetric.

For eg. $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$

Multiplication by scaler

Let $\boldsymbol{A}$ be a mXn matrix and let λ be real number.

Then $\lambda\boldsymbol{A} = K$, where $k_{ij} = \lambda a_{ij}$

Associativity:

$(\lambda\mu)\boldsymbol{C} = \lambda(\mu\boldsymbol{C})$, where $\boldsymbol{C}$ is mXn matrix

$\lambda(\boldsymbol{BC}) = (\lambda\boldsymbol{B})\boldsymbol{C} = \boldsymbol{B}(\lambda\boldsymbol{C}) = (\boldsymbol{BC})\lambda$, where B is mXn matrix and $\boldsymbol{C}$ is a nXp matrix.

$(\lambda\boldsymbol{C})^T = \boldsymbol{C}^T\lambda^T = \boldsymbol{C}^T\lambda = \lambda\boldsymbol{C}^{\boldsymbol{T}}$, where $\lambda = \lambda^T$, real λ

Distributivity:

$(\lambda + \mu)\boldsymbol{C} = \lambda\boldsymbol{C} + \mu\boldsymbol{C}$, where $\boldsymbol{C}$ is a $m \times n$ matrix.

$\lambda(\boldsymbol{B} + \boldsymbol{C}) = \lambda\boldsymbol{B} + \lambda\boldsymbol{C}$, where $\boldsymbol{B}$ & $\boldsymbol{C}$ are $m \times n$ matrices.

Compact representation of system of equations

The system of equations:

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ x_1 - x_2 + 2x_3 &= 2 \\ 2x_1 \qquad + 3x_3 &= 1 \end{aligned}$$

Can be written as:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \\ 2 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

Solving system of linear equations

Ex:

$$\begin{bmatrix} 1 & 0 & 8 & -4 \\ 0 & 1 & 2 & 12 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 42 \\ 8 \end{bmatrix}$$

Particular solution is: (42, 8, 0, 0), since:

$$\begin{bmatrix} 42 \\ 8 \end{bmatrix} = 42 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 8 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Expressing the third column using the first two columns:

$$\begin{bmatrix}8\\2\end{bmatrix} = 8\begin{bmatrix}1\\0\end{bmatrix} + 2\begin{bmatrix}0\\1\end{bmatrix}$$

$$\Leftrightarrow 0 = 8c_1 + 2c_2 - 1c_3 + 0c_4$$

Solution is:

$$(x_1, x_2, x_3, x_4) = (8,2,-1,0)$$

Any scaling of this solution by λ_1 gives 0 vector.

$$\begin{bmatrix}1 & 0 & 8 & -4\\0 & 1 & 2 & 12\end{bmatrix}\left(\lambda_1\begin{bmatrix}8\\2\\-1\\0\end{bmatrix}\right) = \lambda_1(8c_1 + 2c_2 - 1c_3) = 0$$

So, the **general solution** is:

$$\boldsymbol{X} = \begin{bmatrix} 42 \\ 8 \\ 0 \\ 0 \end{bmatrix} + \lambda_1 \begin{bmatrix} 8 \\ 2 \\ -1 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} -4 \\ 12 \\ 0 \\ -1 \end{bmatrix}, \lambda_1, \lambda_2 \; reals$$

Remark. The general approach is:

1. Find a particular solution to $\boldsymbol{Ax} = \boldsymbol{b}$.

2. Find all solutions to $\boldsymbol{Ax} = \boldsymbol{0}$.

3. Combine the solutions from steps 1. and 2. to get the general solution.

Particular and General solutions are not unique.

Elementary Transformations

1. Interchange of two equations (rows in the matrix representing the system of equations).
2. Multiplication of an equation (row) with a non-zero real constant.
3. Addition of two equations (rows).

Ex: For any real number a, find all the solutions:

$$
\begin{aligned}
-2x_1 + 4x_2 - 2x_3 - x_4 + 4x_5 &= -3 \\
4x_1 - 8x_2 + 3x_3 - 3x_4 + x_5 &= 2 \\
x_1 - 2x_2 + x_3 - x_4 + x_5 &= 0 \\
x_1 - 2x_2 \qquad\quad - 3x_4 + 4x_5 &= a
\end{aligned}
$$

Augmented matrix $[A|b]$

$$
\begin{bmatrix}
-2 & 4 & -2 & -1 & 4 & -3 \\
4 & -8 & 3 & -3 & 1 & 2 \\
1 & -2 & 1 & -1 & 1 & 0 \\
1 & -2 & 0 & -3 & 4 & a
\end{bmatrix}
$$

Swapping R1 and R3, gives:

$$\begin{bmatrix} 1 & -2 & 1 & -1 & 1 & 0 \\ 4 & -8 & 3 & -3 & 1 & 2 \\ -2 & 4 & -2 & -1 & 4 & -3 \\ 1 & -2 & 0 & -3 & 4 & a \end{bmatrix}$$

Applying the Elementary Row Operations:

$\mathrm{R2} \longrightarrow \mathrm{R2} - 4\mathrm{R1},\ \mathrm{R3} \longrightarrow \mathrm{R3} + 2\mathrm{R1},\ \mathrm{R4} \longrightarrow \mathrm{R4} - \mathrm{R1}$, gives:

$$\begin{bmatrix} 1 & -2 & 1 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 & -3 & 2 \\ 0 & 0 & 0 & -3 & 6 & -3 \\ 0 & 0 & -1 & -2 & 3 & a \end{bmatrix}$$

Applying $\mathrm{R4} \longrightarrow \mathrm{R4} - \mathrm{R2} - \mathrm{R3}$, gives:

$$\begin{bmatrix} 1 & -2 & 1 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 & -3 & 2 \\ 0 & 0 & 0 & -3 & 6 & -3 \\ 0 & 0 & 0 & 0 & 0 & 1 + a \end{bmatrix}$$

$$\begin{bmatrix} 1 & -2 & 1 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 & -3 & 2 \\ 0 & 0 & 0 & -3 & 6 & -3 \\ 0 & 0 & 0 & 0 & 0 & 1+a \end{bmatrix}$$

Applying $\mathrm{R2} \longrightarrow -\mathrm{R2}, \mathrm{R3} \longrightarrow -\frac{1}{3}\mathrm{R3}$, gives:

$$\begin{bmatrix} 1 & -2 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 3 & -2 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1+a \end{bmatrix}$$

This called **Row Echelon Form**

$$x_1 - 2x_2 + x_3 - x_4 + x_5 = 0$$

$$x_3 - x_4 + 3x_5 = -2$$

$$x_4 - 2x_5 = 1$$

$$0 = a + 1$$

Only for $a = -1$, this system can be solved.

P**articular solution** is: $(2,0,-1,1,0)$

General solution is:

$$\boldsymbol{X} = \begin{bmatrix} 2 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} + \lambda_1 \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2 \\ 0 \\ -1 \\ 2 \\ 1 \end{bmatrix}, \lambda_1, \lambda_2\ reals$$

Def: The leading co-efficient of a row is called **pivot.**

Def: A matrix is in **row-echelon form** if

- All rows that contain only zeros are at the bottom of the matrix; correspondingly, all rows that contain at least one nonzero element are on top of rows that contain only zeros.
- Looking at nonzero rows only, the first nonzero number from the left (also called the **pivot** or the **leading coefficient**) is always strictly to the leading coefficient right of the pivot of the row above it.

Def: The variables corresponding to the pivots in the row-echelon form are called **basic variables** and the other variables are **free variables.**

Def*:* An equation system is in **Reduced Form** *or* **Row-Reduced Echelon Form** or **Row Canonical Form)** if

- It is in row-echelon form.
- Every pivot is 1.
- The pivot is the only nonzero entry in its column.

Exercise: Verify that the following matrix is in the Reduced Row Echelon Form.

$$\boldsymbol{A} = \begin{bmatrix} 1 & 3 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 9 \\ 0 & 0 & 0 & 1 & -4 \end{bmatrix}$$

Ans: All solutions of $\boldsymbol{Ax} = \boldsymbol{b}$, $\boldsymbol{x} \in \mathbb{R}^5$ are:

$$\mathbf{x} \in \mathbb{R}^5 \colon \boldsymbol{x} = \lambda_1 \begin{bmatrix} 3 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 3 \\ 0 \\ 9 \\ -4 \\ -1 \end{bmatrix}, \lambda_1, \lambda_2 \in \mathbb{R}$$

The Minus-1 Trick

A practical trick for reading out the solutions $\boldsymbol{x}$ of a homogeneous system of linear equations $\boldsymbol{Ax} = \mathbf{0}$, where $\boldsymbol{A} \in \mathbb{R}^{k \times n}, x \in \mathbb{R}^n$.

Let $\boldsymbol{A}$ be in reduced row-echelon form without any rows that just contain zeros, i.e.,

$$\boldsymbol{A} = \begin{bmatrix} 0 & \cdots & 0 & 1 & * & \cdots & * & 0 & * & \cdots & * & 0 & * & \cdots & * \\ \vdots & & \vdots & 0 & 0 & \cdots & 0 & 1 & * & \cdots & * & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & 0 & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & 0 & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & * & \cdots & * \end{bmatrix}$$

where * are arbitrary real number, and first nonzero entry per row must be 1 and all other entries in corresponding column must be 0.

Columns with the pivots, are standard unit vectors $\boldsymbol{e_1}, \boldsymbol{e}_2, \dots\ \boldsymbol{e}_k$ in $\mathbb{R}^k$.

Extend this matrix to an $n \times n$ matrix $\widetilde{\boldsymbol{A}}$ by adding $n-k$ rows of the form $[0, \dots 0, -1,0, \dots 0]$ so that the diagonal of the augmented matrix $\widetilde{\boldsymbol{A}}$ contains either 1 or -1.

Then the columns of $\widetilde{\boldsymbol{A}}$ that contain the -1 as pivots are the solutions of homogeneous system $\boldsymbol{Ax} = \boldsymbol{b}$. These columns form the basis of solution space of $\boldsymbol{Ax} = \boldsymbol{b}$ and are called **kernel or null space.**

Ex: Consider this matrix already in REF:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 3 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 9 \\ 0 & 0 & 0 & 1 & -4 \end{bmatrix}$$

Augmented 5×5 matrix (by adding rows of the form $[0, \ldots 0, -1, 0, \ldots 0]$ at the places where the pivots on the diagonal are missing) is:

$$\widetilde{\boldsymbol{A}} = \begin{bmatrix} 1 & 3 & 0 & 0 & 3 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 9 \\ 0 & 0 & 0 & 1 & -4 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix}$$

From $\widetilde{\boldsymbol{A}}$, we can easily read the solution from the columns which have -1 in the diagonal.

$$\mathbf{x} = \lambda_1 \begin{bmatrix} 3 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 3 \\ 0 \\ 9 \\ -4 \\ -1 \end{bmatrix}, \lambda_1, \lambda_2 \; reals$$

Calculating the Inverse

To compute inverse $\boldsymbol{A}^{-1}$ of a nXn matrix $\boldsymbol{A}$.

To find a matrix $\boldsymbol{X}$ that satisfies $\boldsymbol{AX} = \boldsymbol{I}_n$.

Then $\boldsymbol{X} = \boldsymbol{A}^{-1}$.

Writing this as a set of simultaneous linear equations $\boldsymbol{AX} = \boldsymbol{I}_n$ where we solve for $\boldsymbol{X} = [x_1| \dots |x_n]$

Use the augmented matrix notation for a compact representation of this set of systems of linear equations and obtain $[\boldsymbol{A}|\boldsymbol{I}_n] \dots \dots \{\boldsymbol{I}_n|\boldsymbol{A}]$

So, if we bring the augmented equation system into reduced row-echelon form, we can read out the inverse on the right-hand side of the equation system.

Hence, determining the inverse of a matrix is equivalent to solving systems of linear equations.

Ex: Find the inverse of

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Augmented matrix is:

$$\left[\begin{array}{cccc|cccc} 1 & 0 & 2 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{array}\right]$$

The Row Echolen Form is:

$$\left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & -1 & 2 & -2 & 2 \\ 0 & 1 & 0 & 0 & 1 & -1 & 2 & -2 \\ 0 & 0 & 1 & 0 & 1 & -1 & 1 & -1 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 & 2 \end{array}\right]$$

Algorithms for Solving a System of Linear Equations

Given a system of linear equations of the type $\boldsymbol{Ax = b}$.

$$\Leftrightarrow \boldsymbol{A^T A x = A^T b}$$

$$\Leftrightarrow \boldsymbol{x = (A^T A)^{-1} A^T b}$$

Gaussian elimination

Gaus-Seidel method

Ex:

Solve the following system of equations:

$2w+4x+5y-4z=10$

$5w+x+2y+6z=12$

$3w+5x+7y-2z=15$

$5w+2x-2y+3z=14$

Ans: $w = 65/64 = 1.015625$

$x = 197/64 = 3.47893$

$y = -9/32 = -0.28125$

$z = 47/64 = 0.734375$

Set of matrices as Vector spaces

- Vector spaces
- Vector subspaces
- Linearly independent vectors

Vector spaces

Def: A real-valued **vector space** V = $(\mathcal{V},\oplus,\otimes)$ is a set $\mathcal{V}$ with two operations:

$\oplus : \mathcal{V} X \mathcal{V} \longrightarrow \mathcal{V}$ and $\otimes : \mathcal{R} X \mathcal{V} \longrightarrow \mathcal{V}$, Where:

1. $(\mathcal{V},\oplus)$ is Abelian group.
2. Distributivity:

   (i) $\forall\ \lambda \in \mathcal{R}, \boldsymbol{x}, \boldsymbol{y} \ \in \mathcal{V}$

   $$\Rightarrow \lambda \otimes (\boldsymbol{x} \oplus \boldsymbol{y}) = \lambda \otimes \boldsymbol{x} \oplus \lambda \otimes \boldsymbol{y}$$

   (ii) $\forall\ \lambda, \mu \in \mathcal{R},\ \boldsymbol{x} \in \mathcal{V}$

   $$\Rightarrow (\lambda \oplus \mu) \otimes \boldsymbol{x} = \lambda \otimes \boldsymbol{x} \oplus \mu \otimes \boldsymbol{x}$$

3. Associativity:

$\forall\ \lambda, \mu \in \mathcal{R}, \boldsymbol{x} \in \mathcal{V} \Rightarrow \lambda \otimes (\mu \otimes \boldsymbol{x}) = (\lambda \otimes \mu) \otimes \boldsymbol{x}$

4. Neutral element w.r.t. $\otimes$

$$\forall\ \boldsymbol{x} \in \mathcal{V} \Rightarrow 1 \otimes \boldsymbol{x} = \boldsymbol{x}$$

$\boldsymbol{x} \in \mathcal{V}$ are called **vectors.**

The neutral element of $(V, \oplus)$ is **zero vector** $0 = [0, \ldots 0]$.

Inner operation $\oplus$ is called **vector addition.**

$\lambda \in \mathcal{R}$ are called the **scalers.**

Outer operation $\otimes$ is **multiplication by scalers.**

Ex 1: $\mathcal{V} = \mathbb{R}^{m\times n}, \mathrm{m}, \mathrm{n} \in \mathbb{N}$, is a vector space with:

Addition $\boldsymbol{A} \oplus \boldsymbol{B} = \begin{bmatrix} a_{11}+b_{11} & \cdots & a_{1n}+b_{1n} \\ \vdots & & \vdots \\ a_{m1}+b_{m1} & \cdots & a_{mn}+b_{mn} \end{bmatrix}$

for all $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{V}$.

Scaler multiplication: $\lambda \otimes \boldsymbol{A} = \begin{bmatrix} \lambda a_{11} & \cdots & \lambda a_{1n} \\ \vdots & & \vdots \\ \lambda a_{m1} & \cdots & \lambda a_{mn} \end{bmatrix}$

Def: Let V be a vector space and let $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_k \in V$ be a set of finite number of vectors in V. Then every $\boldsymbol{v} \in V$ of the form:

$$\boldsymbol{v} = \lambda_1 \boldsymbol{x}_1 + \lambda_2 \boldsymbol{x}_2 + \ \dots + \lambda_k \boldsymbol{x}_k = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i \in V$$

with $\lambda_1, \lambda_2, \dots \lambda_k \in \mathbb{R}$ is a **<u>linear combination</u>** of the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_k$

Ex: $\begin{bmatrix} 6 \\ 7 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 3 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + 4 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Consider a vector space V with $k \in \mathcal{N}$ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_k \in V$. If there is a non-trivial linear combination, such that $0 = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i$ with at least one $\lambda_i \neq 0$, $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_k$ are **linearly dependent.**

If only trivial solution exists,

i.e., $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 0, \boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_k$ are **linearly independent.**

Ex: The vectors $\begin{bmatrix} 1 \\ 0 \end{bmatrix} and \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ in $\mathbb{R}^2$ are linearly independent.

To find whether vectors are linearly independent or not:

1. k vectors are either linearly dependent or linearly independent.
2. If at least one of the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_k$ is 0 then they are linearly dependent.
3. If two vectors are identical then linearly dependent.
4. The vectors $\{\ \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_k : \boldsymbol{x}_i \neq 0, i = 1, \ldots k\}$, k $\geq$ 2, are linearly dependent if and only if (at least) one of them is a linear combination of the others. In particular, if one vector is a multiple of another vector, i.e., $\boldsymbol{x}_i = \lambda \boldsymbol{x}_j\ , \lambda \in \mathbb{R}$, then $\{\boldsymbol{x}_1, \ldots\ \boldsymbol{x}_k : \boldsymbol{x}_i \neq 0, i = 1, \ldots k\}$ is linearly dependent.

4. A practical way of checking whether vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_k \in V$ are linearly independent is to use Gaussian elimination:

Write all vectors as columns of a matrix A and perform Gaussian elimination until matrix is in row echelon form.

The pivot columns indicate the vectors, which are linearly independent of the vectors on the left. Note that there is an ordering of vectors when matrix is built.

– The non-pivot columns can be expressed as linear combinations of the pivot columns on their left.

For example: In the row-echelon form:

$$\begin{bmatrix} 1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

1st and 3rd columns are pivot columns.

2nd column is a non-pivot column (because it is three times the first column.)

All column vectors are linearly independent if and only if all columns are pivot columns. If there is at least one non-pivot column, the columns (and, therefore, the corresponding vectors) are linearly dependent.

Ex: In R^4, find if the following vectors are linearly independent.

$$\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 2 \\ -3 \\ 4 \end{bmatrix}, \boldsymbol{x}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 2 \end{bmatrix} \text{ and } \boldsymbol{x}_3 = \begin{bmatrix} -1 \\ -2 \\ 1 \\ 1 \end{bmatrix}$$

Solving for $\lambda_1, \lambda_2, \lambda_3$:

$$\lambda_1\boldsymbol{x}_1 + \lambda_2\boldsymbol{x}_2 + \lambda_3\boldsymbol{x}_3 = \lambda_1 \begin{bmatrix} 1 \\ 2 \\ -3 \\ 4 \end{bmatrix} + \lambda_2 \begin{bmatrix} 1 \\ 1 \\ 0 \\ 2 \end{bmatrix} + \lambda_3 \begin{bmatrix} -1 \\ -2 \\ 1 \\ 1 \end{bmatrix} = 0$$

Apply Elementary Row Operations on $\begin{bmatrix} 1 & 1 & -1 \\ 2 & 1 & -2 \\ -3 & 0 & 1 \\ 4 & 2 & 1 \end{bmatrix}$

We get: $\begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$

Here, every column of the matrix is a pivot column. Therefore, there is no non-trivial solution, and we require $\lambda_1 = 0, \lambda_2 = 0, \lambda_3 = 0$

to solve the equation system.

Hence, vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ are linearly independent.

Note:

In a vector space $V,\ m$ linear combinations of k vectors $x_1, x_2, \dots\ x_k$ are linearly dependent if $m\ >\ k$.

Check: Home work

Basis and Rank of matrices

- Basis of a vector space
- Rank of a matrix

Generating Set and Basis

Def: Consider a vector space $\mathrm{V} = (\mathcal{V}, \oplus, \otimes)$ and set of vectors $\mathcal{A} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_k\} \sqsubseteq V$. If every vector $\boldsymbol{v} \in \mathcal{V}$ can be expressed as a linear combination of $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_k$, then $\mathcal{A}$ is called a **generating set of V .**

The set of all linear combinations of vectors in $\mathcal{A}$ is called the **span of $\mathcal{A}$**

If $\mathcal{A}$ spans the vector space $\mathcal{V}$, we write:

$\mathcal{V} = \text{span}[\mathcal{A}]$ or

$\mathcal{V} = \text{span}[\boldsymbol{x}_1, \dots \boldsymbol{x}_k]$ or

$\mathcal{V} = \text{Linearspan}[\mathcal{A}]$

Def: Consider a vector space $\mathrm{V} = (\mathcal{V}, \oplus, \otimes)$ and $\mathcal{A} \sqsubseteq \mathcal{V}$. A generating set $\mathcal{A}$ of V is called **minimal generating set** if there exists no smaller set $\tilde{\mathcal{A}} \sqsubseteq \mathcal{A} \sqsubseteq \mathcal{V}$ that spans V.

Every linearly independent generating set of V is minimal and is called a **basis of V**

Let $\mathrm{V} = (\mathcal{V}, \oplus, \otimes)$ be a vector space and $\mathcal{B} \sqsubseteq \mathcal{V}$. Then, the following statements are equivalent:

- $\mathcal{B}$ is a basis of V.
- $\mathcal{B}$ is a minimal generating set.
- $\mathcal{B}$ is a maximal linearly independent set of vectors in V , i.e., adding any other vector to this set will make it linearly dependent.
- Every vector $\boldsymbol{x} \in \mathrm{V}$ is a linear combination of vectors from $\mathcal{B}$, and every linear combination is unique, i.e., with $\boldsymbol{x} = \sum_{i=1}^{k} \lambda_i \boldsymbol{b}_i = \sum_{i=1}^{k} \mu_i b_i$, and $\lambda_i, \mu_i \in \mathbb{R}, \boldsymbol{b}_i \in \mathcal{B}$, it implies $\lambda_i = \mu_i, i = 1, 2, \dots k$

Ex: In $\mathbb{R}^3$, the **canonical/standard basis** is

$$\mathcal{B} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

Ex: Different bases in $\mathbb{R}^3$ are:

$$\mathcal{B}_1 = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}, \mathcal{B}_2 = \left\{ \begin{bmatrix} 0.5 \\ 0.8 \\ 0.4 \end{bmatrix}, \begin{bmatrix} 1.8 \\ 0.3 \\ 0.3 \end{bmatrix}, \begin{bmatrix} -2.2 \\ -1.3 \\ 3.5 \end{bmatrix} \right\}$$

Basis of a vector space need not be unique although no. of elements in a basis is unique.

Ex: The set

$$\boldsymbol{A} = \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ -4 \end{bmatrix} \right\}$$

is linearly independent, but not a generating set (and no basis) of $\mathbb{R}^4$.

e.g. the vector $\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ cannot be obtained by a linear combination of elements in $\boldsymbol{A}$.

Ex : In the vector space of square matrix over real numbers of dimension 4, the basis is:

$$B = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}.$$

(i) They are linearly independent

(ii) Any 2X2 matrix can be written as:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ c & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & d \end{bmatrix}$$

Let $A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$ where $a_{ij} \in R$.

1. The row space of A is given by linearspan$[(a_{11}, \dots a_{1n}), \dots (a_{m1}, \dots\ a_{mn})]$ is a subspace of R^n.
2. The column space of A is given by linearspan$[(a_{11}, \dots a_{m1}), \dots (a_{1n}, \dots\ a_{mn})]$ is a subspace of R^m.
3. The set N(A)={$\boldsymbol{x} \in R^n | \boldsymbol{Ax} = \boldsymbol{0}$ } is said to be the **null space** of A.
4. The set R(A)={$\boldsymbol{b} \in R^m | \boldsymbol{Ax} = \boldsymbol{b}$ } is said to be the **range of A.**

Ex 1: Dim $(\mathbb{R}^2) = 2$ over the set of real numbers.

Ex 2: Dim $(\mathbb{R}^3) = 3$ over the set of real numbers.

Ex 3: Dimension of the vector space of $m \times n$-matrix over real numbers is mn.

Ex 4: Dimension of vector space of polynomials of degree n over real numbers is $n + 1$.

Ex 5: Dimension of vector space of all polynomials is ∞.

Theorem: Let $B = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \dots\ \boldsymbol{b}_n\}$ be a Basis of n −dimensional vector space V over the set of real numbers.

If S is a subset of vectors of V having more than n vectors, then S is linearly dependent.

If S is a subset of vectors of V having less than n vectors, then S can be extended to a basis of V.

Theorem: Let V be a finite dimensional vector space over the set of real numbers. Then any two basis of V must have the same number of vectors.

Ex: In $\mathbb{R}^3$, over the field of real numbers, then the standard basis is $\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}$.

Another basis is : $\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}1\\1\\1\end{bmatrix}$.

Rank

Def: The number of linearly independent columns of a matrix $A \in \mathbb{R}^{m\times n}$ equals the number of linearly independent rows and is called the **rank of A** and is denoted by rk(**A**).

Properties of rank

1. $rk(\boldsymbol{A}) = rk(\boldsymbol{A}^T)$, i.e. column rank=row rank.
2. The columns of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ span a subspace $U \sqsubseteq \mathbb{R}^m$ with $\dim(U) = rk(\boldsymbol{A})$.

   This subspace is called the ***image or range***.

A basis of U can be found by applying Gaussian elimination to $\boldsymbol{A}$ to identify the pivot columns.

3. Rows of $A \in \mathbb{R}^{m\times n}$ span a subspace $W \sqsubseteq \mathbb{R}^n$ with $\dim(W) = rk(\boldsymbol{A})$. A basis of W can be found by applying Gaussian elimination to $\boldsymbol{A}^T$.

4. For all $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ holds that $\boldsymbol{A}$ is regular (invertible) if and only if $rk(\boldsymbol{A}) = n$.

5. For all $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and all $\boldsymbol{b} \in \mathbb{R}^m$, it holds that linear equation $\boldsymbol{A}x = \boldsymbol{b}$ can be solved if and only if $rk(\boldsymbol{A}) = rk(\boldsymbol{A}|\boldsymbol{b})$, where $\boldsymbol{A}|\boldsymbol{b}$ denotes the augmented system.

6. For $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ the subspace of solutions for $\boldsymbol{Ax} = \boldsymbol{0}$ possesses dimension $n - rk(\boldsymbol{A})$.

   This subspace is called **kernel or the null space.**

7. A matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has full rank if its rank equals the largest possible rank for a matrix of the same dimensions. This means that the rank of a full-rank matrix is the lesser of the number of rows and columns, i.e., $rk(\boldsymbol{A}) = \min(m, n)$.

Def: A matrix is said to be **rank deficient** if it does not have full rank.

Ex:

1. **A** = $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ has two linearly independent rows/ columns. So rk(**A**) = 2.

2. To find the rank of **A** = $\begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix}$, apply Gaussian Elimination to get $\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix}$. Number of linearly independent rows and columns is 2. So rk(**A**)=2.

Ex: Find the rank of the following matrix

$$A = \begin{bmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{bmatrix}$$

Ans: Rank is 1

Linear Mappings/
Linear Transformations

Applications:

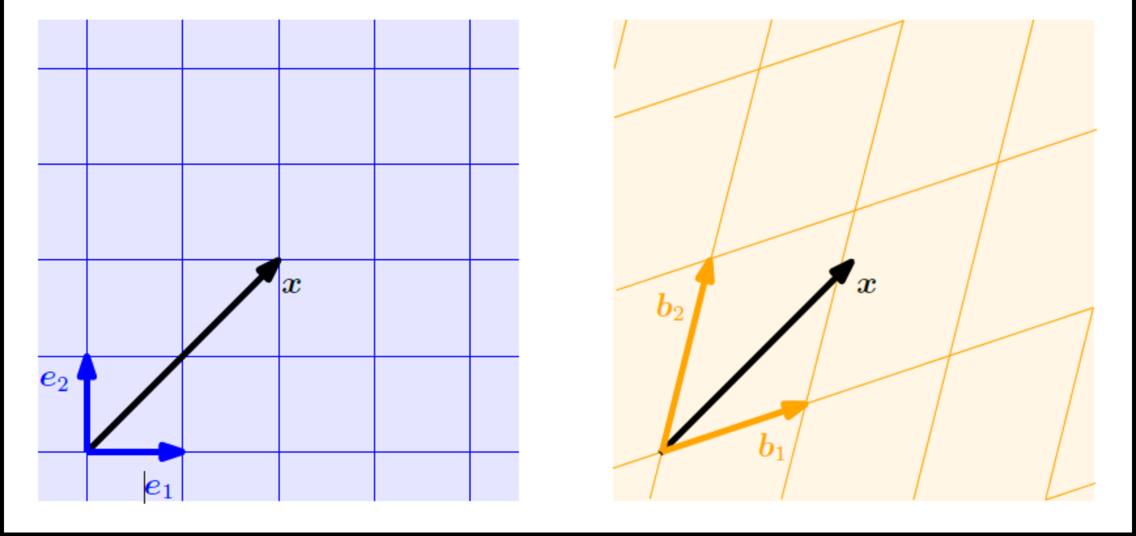
Derivatives, In 2D and 3D animation, Rotations and scaling, projection from one space to another.

Two different coordinate systems defined by two sets of basis vectors.

A vector $\boldsymbol{x}$ has different coordinate representations depending on which coordinate system is chosen.

Coordinates of $\boldsymbol{x}$ w.r.t. standard basis (e_1, e_2) of R^2 is $[2,2]^T$and w.r.t. $(\boldsymbol{b}_1, \boldsymbol{b}_2)$ it is $[1.09,0.72]^T$

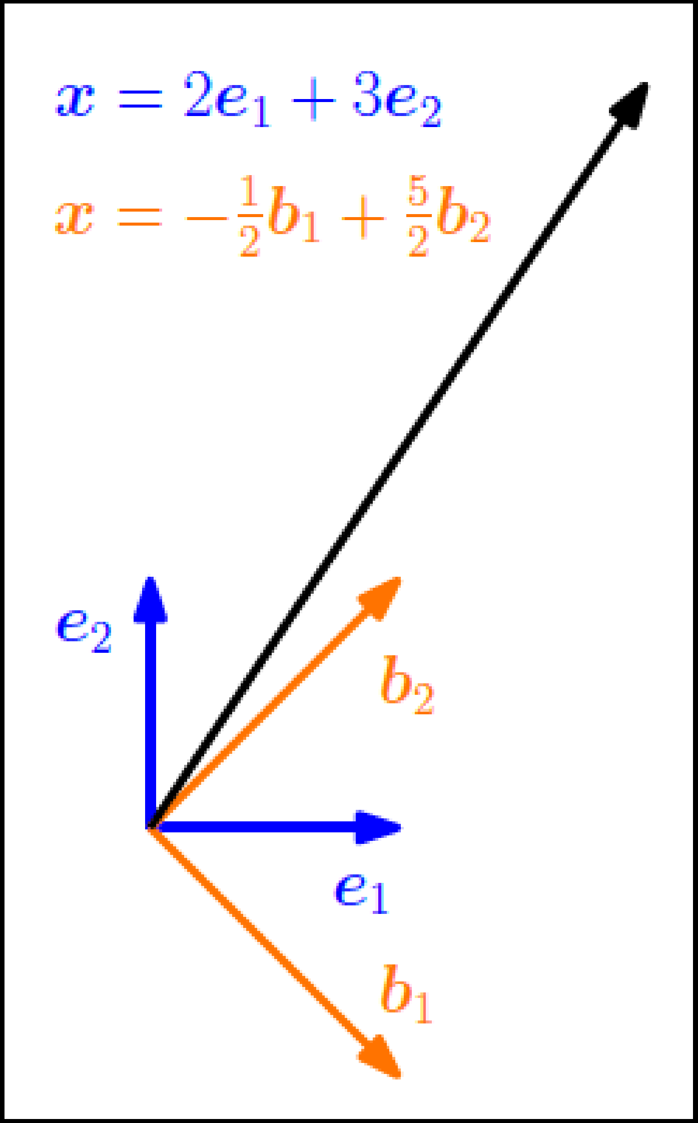
i.e. $\boldsymbol{x} = 1.09\boldsymbol{b}_1 + 0.72\boldsymbol{b}_2$



For $\boldsymbol{x} \in \mathbb{R}^2$ with coordinates $[2,3]^T$ w.r.t. standard basis (e_1, e_2).

So, $\boldsymbol{x} = 2e_1 + 3e_2$.

If we use basis vectors $\boldsymbol{b}_1 = [1,-1]^T$ & $\boldsymbol{b}_2 = [1,1]^T$ we will obtain the coordinates $\frac{1}{2}[-1,5]^T$ to represent the same vector w.r.t. $(\boldsymbol{b}_1, \boldsymbol{b}_2)$.

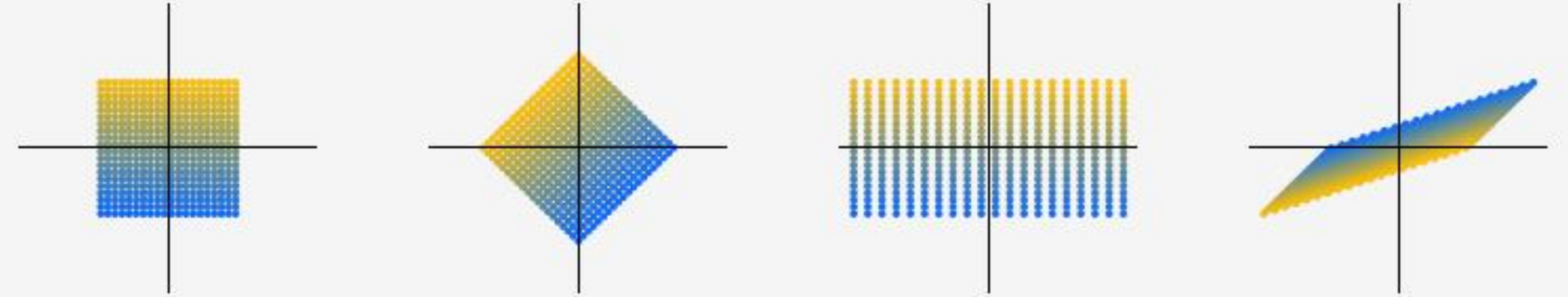


Examples of linear transformations of vectors shown as dots in (a)

(b) Rotation by $45°$

(c) Stretching of the horizontal coordinates by 2

(d) Combination of reflection, rotation and stretching.



(a) Original data.

(b) Rotation by $45°$.

(c) Stretch along the horizontal axis.

(d) General linear mapping.

These linear transformations (in fig above) are of a set of vectors in $\mathbb{R}^2$ with transformation matrices:

$$\boldsymbol{A}_1 = \begin{bmatrix} \cos({}^{\pi}/_{4}) & -\sin({}^{\pi}/_{4}) \\ \sin({}^{\pi}/_{4}) & \cos({}^{\pi}/_{4}) \end{bmatrix}$$

$$\boldsymbol{A}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\boldsymbol{A}_3 = \frac{1}{2}\begin{bmatrix} 3 & -1 \\ 1 & -1 \end{bmatrix}$$

Def: Two matrices $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{n}}$ are **equivalent** if there exist regular matrices $\boldsymbol{S} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{n}}$ and $\boldsymbol{T} \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{m}}$, such that $\widetilde{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{S}$.

Def: Two matrices $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{n}}$ are **similar** if there exist regular matrices $\boldsymbol{S} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{n}}$ with $\widetilde{\boldsymbol{A}} = \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$.

Remark. Similar matrices are always equivalent. However, equivalent matrices are not necessarily similar.

Def: A symmetric matrix $A \in \mathbb{R}^{nn\times}$ is said to be **symmetric positive definite or positive definite** if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0 \;\; \forall\, \boldsymbol{x} \in V\backslash\{\mathbf{0}\}$.

Further, If $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0 \;\forall\, \boldsymbol{x} \in V\backslash\{\mathbf{0}\}$, then $\boldsymbol{A}$ is called **symmetric positive semidefinite.**

Ex 1: Let $\boldsymbol{A} = \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix}$. This is symmetric matrix.

And $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$= 9x_1^2 + 12x_1x_2 + 5x_2^2 = (3x_1 + 2x_2)^2 + x_2^2$$

This is > 0 or all nonzero **x**. So $\boldsymbol{A}$ is symmetric positive definite.

Ex 2: $\boldsymbol{B} = \begin{bmatrix} 9 & 6 \\ 6 & 3 \end{bmatrix}$ is symmetric but not positive definite, as $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = (3x_1 + 2x_2)^2 - x_2^2$ may be < 0.

Matrix Decomposition

Three aspects of matrices:

1. How to summarize matrices,
2. how to decompose matrices, and
3. how these decompositions can be used for matrix approximations.

Consider square matrix of dimension n over the field of real numbers

Let $\boldsymbol{A}$ be a square matrix of order n over the set of real numbers.

Then , **determinant** of $\boldsymbol{A}$ is found as follows:

If n=1, then det($\boldsymbol{A}$)=a_{11}

If n=2, then det($\boldsymbol{A}$)=$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$

If n=3, then det($\boldsymbol{A}$)=$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$

$$= a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

…..

Ex: Find the det($\boldsymbol{A}$) where

$$\boldsymbol{A} = \begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix} = 1\begin{vmatrix} 5 & 6 \\ 8 & 9 \end{vmatrix} - 2\begin{vmatrix} 4 & 6 \\ 7 & 9 \end{vmatrix} + 3\begin{vmatrix} 4 & 5 \\ 7 & 8 \end{vmatrix}$$

$$= (-3) - 2(-6) + 3(-3)$$

$$= 0$$

Def: Matrix having determinant equal to zero are called **singular matrix.**

Theorem: Any square matrix $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{n}}$ is invertible if and only if det($\boldsymbol{A}$) $\neq 0$.

Def: For a (**lower-triangular or upper triangular**) matrix **T** of order n over real numbers, the determinant is the product of its diagonal elements. That is:

$$\det(\boldsymbol{T}) = \prod_{i=1}^{n} a_{ii}$$

Ex: If $\boldsymbol{T} = \begin{vmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{vmatrix}$, then $\det(\boldsymbol{T}) = 24.$

Theorem: **(Laplace Expansion).** Consider a square matrix $\boldsymbol{A}$ of order n over reals. Then, for all $j = 1, 2, \dots n$:

1. Expansion along column j

$$\det(\boldsymbol{A}) = \sum_{k=1}^{n} (-1)^{k+j}\, a_{kj} \det(\boldsymbol{A}_{k,j})$$

1. Expansion along row j

$$\det(\boldsymbol{A}) = \sum_{k=1}^{n} (-1)^{k+j}\, a_{jk} \det(\boldsymbol{A}_{j,k})$$

Where $\boldsymbol{A}_{k,j}$ is a sub-matrix of $\boldsymbol{A}$ of order (n-1) obtained by deleting k^{th} row and j^{th} column.

Ex: Compute determinant of $\boldsymbol{A} = \begin{vmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 0 & 0 & 1 \end{vmatrix}$

using the Laplace expansion along the first row, Applying expansion along row 1, gives:

$$\begin{vmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 0 & 0 & 1 \end{vmatrix}$$

$$= (-1)^{1+1}.1 \begin{vmatrix} 1 & 2 \\ 0 & 1 \end{vmatrix} + (-1)^{1+2}.2 \begin{vmatrix} 3 & 2 \\ 0 & 1 \end{vmatrix} + (-1)^{1+3}.3 \begin{vmatrix} 3 & 1 \\ 0 & 0 \end{vmatrix}$$

$$= 1(1-0) - 2(3-0) + 3(0-0)$$

$$= -5.$$

Theorem: $\mathbf{A} \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$ has $\det(\boldsymbol{A}) \neq 0$ if and only if $rk(\boldsymbol{A}) = n$. In other words, **A** is invertible if and only if it is of full rank.

Def: The **trace** of a square matrix $\mathbf{A} \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$ is sum of the diagonal elements of $\boldsymbol{A}$, i.e. $tr(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}$

Properties of trace:

1. $tr(\boldsymbol{A} + \boldsymbol{B}) = tr(\boldsymbol{A}) + tr(\boldsymbol{B})$ for all $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$
2. $tr(\alpha\boldsymbol{A}) = \alpha . tr(\boldsymbol{A}), \alpha \in \mathbb{R}$, for $\mathbf{A} \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$
3. $tr(\boldsymbol{I}_n) = n$
4. $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$ for $\boldsymbol{A} \in \mathbb{R}^{\mathrm{n}\times k}, \boldsymbol{B} \in \mathbb{R}^{k\times n}$

Def: Let $\mathbf{A} \in \mathbb{R}^{\mathrm{n \times n}}$, and $\lambda \in \mathbb{R}$,

$$p_A^{\lambda} = \det(\boldsymbol{A} - \lambda \boldsymbol{I})$$
$$= c_0 + c_1\lambda + c_2\lambda^2 + \ \ldots + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n$$

For $c_0, c_1, \ldots\ c_{n-1} \in \mathbb{R}$, is the **characteristic polynomial** of $\boldsymbol{A}$.

In particular:

$$c_0 = \det(\boldsymbol{A})$$
$$c_{n-1} = (-1)^{n-1} tr(\boldsymbol{A})$$

Def: Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a square matrix, Then $\lambda \in \mathbb{R}$ is an **eigenvalue** of $\boldsymbol{A}$ and $\boldsymbol{x} \in \mathbb{R}^n\backslash\{0\}$ is the corresponding **eigenvector** of $\boldsymbol{A}$ if $\boldsymbol{Ax} = \lambda\boldsymbol{x}$.

Following statements are equivalent:

1. $\lambda \in \mathbb{R}$ is an eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

2. There exists an $\boldsymbol{x} \in \mathbb{R}^n\backslash\{0\}$ with $\boldsymbol{Ax} = \lambda\boldsymbol{x}$, *or* equivalently $(\boldsymbol{A} - \lambda\boldsymbol{I_n})x = 0$ can be solved non-trivially, i.e. $\boldsymbol{x} \neq 0$.

3. $rk(\boldsymbol{A} - \lambda\boldsymbol{I}_n) < n$.

4. $det(\boldsymbol{A} - \lambda\boldsymbol{I}_n) = 0$.

Ex: Find the eigenvalues and eigenvectors of the 2X2 matrix $\boldsymbol{A} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}$.

Characteristic polynomial:

$$p_{\boldsymbol{A}}(\lambda) = \det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$$

$$\det\left(\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - \lambda\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = \begin{vmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix}$$

$$= (4-\lambda)(3-\lambda) - 2$$

$$= 10 - 7\lambda + \lambda^2$$

$$(2-\lambda)(5-\lambda) = 0$$

Gives $\lambda_1 = 2$ and $\lambda_2 = 5$

Eigenvectors and Eigenspaces

$$\begin{vmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix} x = 0$$

For $\lambda = 5$,

$$\begin{vmatrix} 4-5 & 2 \\ 1 & 3-5 \end{vmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Leftrightarrow \begin{vmatrix} -1 & 2 \\ 1 & -2 \end{vmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Solving gives: $E_5 = span[\begin{bmatrix} 2 \\ 5 \end{bmatrix}]$

This eigenspace is one-dimensional as it possesses a single basis vector.

Similarly for $\lambda = 2$,

$$\begin{vmatrix} 4-2 & 2 \\ 1 & 3-2 \end{vmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Leftrightarrow \begin{vmatrix} 2 & 2 \\ 1 & 1 \end{vmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This means, any vector $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ where $x_1 = x_2$ is a solution.

So $\qquad E_2 = span[\begin{bmatrix} 1 \\ -1 \end{bmatrix}]$

Note: In general, there may be multiple identical eigenvalues and the eigenspace may have more than one dimension.

Graphical Intuition in Two Dimensions

Intuition for determinants, eigenvectors, and eigenvalues using 5 different linear mappings on a square grid of points, centered at origin:

$$(1)\ \mathbf{A}_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix}$$

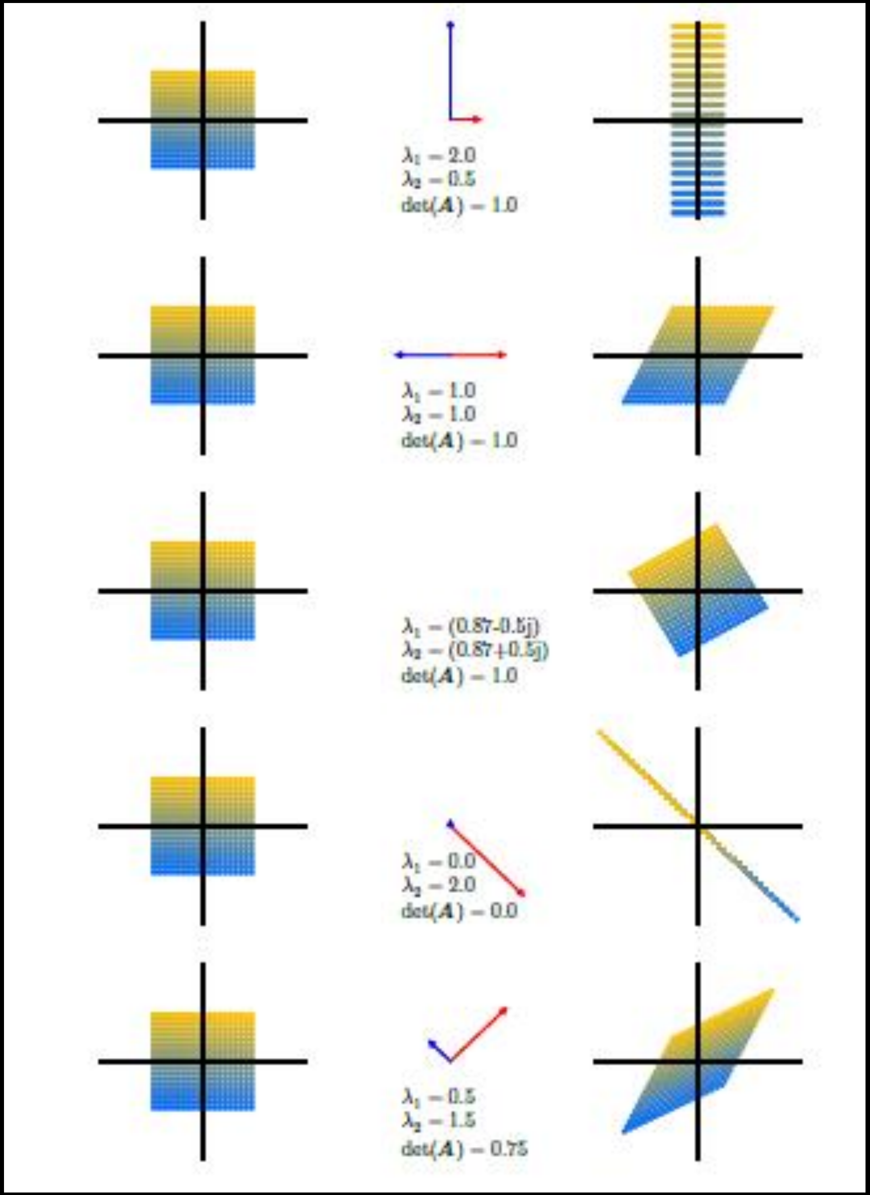
$$(2)\ \mathbf{A}_2 = \begin{bmatrix} 1 & 1/2 \\ 0 & 1 \end{bmatrix}$$

$$(3)\ \mathbf{A}_3 = \begin{bmatrix} \cos\left(\frac{\pi}{6}\right) & -\sin\left(\frac{\pi}{6}\right) \\ \sin\left(\frac{\pi}{6}\right) & \cos\left(\frac{\pi}{6}\right) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \sqrt{3} & -1 \\ 1 & \sqrt{3} \end{bmatrix}$$

$$(4)\ \mathbf{A}_4 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

$$(5)\ \mathbf{A}_5 = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$$

Five linear mappings and their associated transformation matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$, $\boldsymbol{A}_4$, $\boldsymbol{A}_5$ projecting 400 color-coded points $\boldsymbol{x} \in \mathbb{R}^2$ (left column) onto target points $\boldsymbol{A}_i\boldsymbol{x}$ Aix (right column). The central column depicts the first eigenvector, stretched by its associated eigenvalue λ_1, and second eigenvector stretched by its eigenvalue λ_2. Each row depicts the effect of one of five transformation matrices $\boldsymbol{A}_i$ with respect to the standard basis.


$\lambda_1 = 2.0$
$\lambda_2 = 0.5$
$\det(\boldsymbol{A}) = 1.0$
$\lambda_1 = 1.0$
$\lambda_2 = 1.0$
$\det(\boldsymbol{A}) = 1.0$
$\lambda_1 = (0.87\text{-}0.5j)$
$\lambda_2 = (0.87+0.5j)$
$\det(\boldsymbol{A}) = 1.0$
$\lambda_1 = 0.0$
$\lambda_2 = 2.0$
$\det(\boldsymbol{A}) = 0.0$
$\lambda_1 = 0.5$
$\lambda_2 = 1.5$
$\det(\boldsymbol{A}) = 0.75$

Ex: Let $\boldsymbol{A} = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$.

$$p_{\boldsymbol{A}}(\lambda) = -(\lambda - 1)^2(\lambda - 7)$$

Eigon values are:

$\lambda_1 = 1$ (repeated eigen value)and $\lambda_2 = 7$.

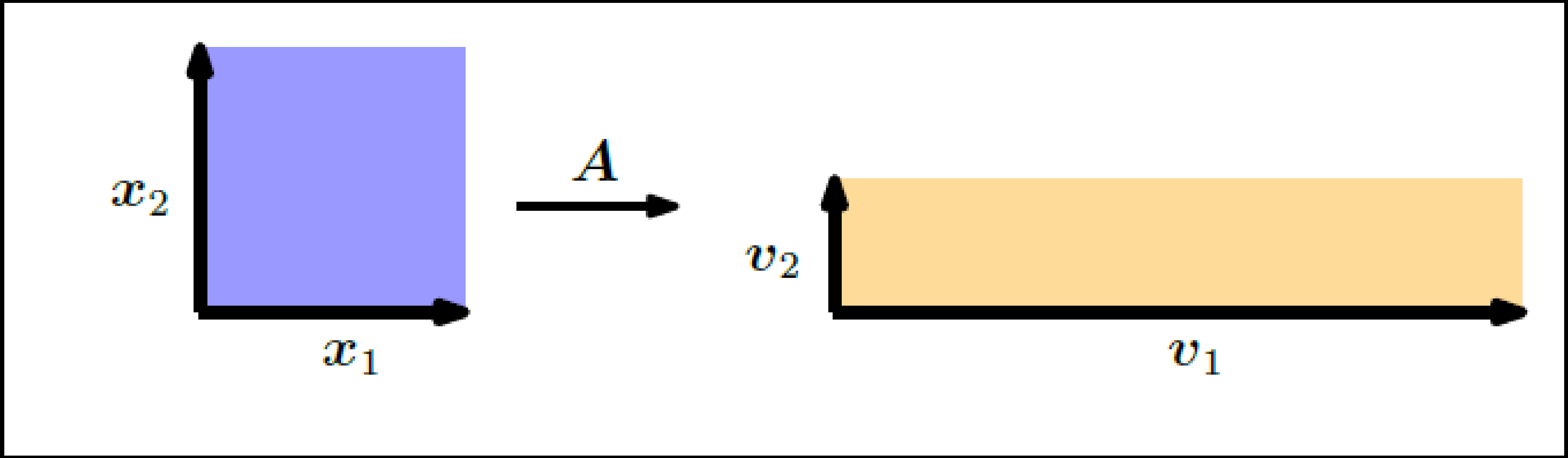
Eigen spaces are:

$E_1 = span[\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}]$ for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$

$E_7 = span[\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}]$ for $\boldsymbol{x}_3$

Geometric interpretation of eigenvalues.

The eigenvectors of $\boldsymbol{A}$ get stretched by the corresponding eigenvalues. The area of the unit square changes by $|\lambda_1\lambda_2|$, the perimeter changes by a factor of $\frac{1}{2}(|\lambda_1| + |\lambda_2|)$



Ex: (Google's PageRank – Webpages as Eigenvectors)

Google uses the eigenvector corresponding to the maximal eigenvalue of a matrix $\boldsymbol{A}$ to determine the rank of a page for search. The idea for the PageRank algorithm, developed at Stanford University by Larry Page and Sergey Brin in 1996, was that the importance of any web page can be approximated by the importance of pages that link to it. For this, they write down all web sites as a huge directed graph that shows which page links to which. PageRank computes the weight (importance) $x_i \geq 0$ of a web site a_i by counting the number of pages pointing to a_i.

Moreover, PageRank takes into account the importance of the web sites that link to a_i. The navigation behavior of a user is then modeled by a transition matrix $\boldsymbol{A}$ of this graph that tells us with what (click) probability somebody will end up on a different web site. The matrix $\boldsymbol{A}$ has the property that for any initial rank/importance vector $\boldsymbol{x}$ of a web site the sequence $\boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}^2\boldsymbol{x}, \ldots$ converges to a vector $\boldsymbol{x}^*$. This vector is called the PageRank and satisfies $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{x}^*$, i.e., it is an eigenvector (with corresponding eigenvalue 1) of $\boldsymbol{A}$. After normalizing $\boldsymbol{x}^*$, such that $||\boldsymbol{x}^*|| = 1$, we can interpret the entries as probabilities.

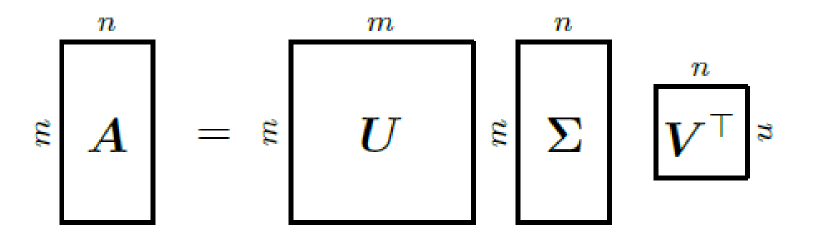
Singular Value Decomposition

It can be applied to any rectangular matrix.

It always exists.

It quantifies the change between the underlying geometry of two vector spaces, where $\boldsymbol{A}$ represent a linear transformation $\phi: V \longrightarrow W$.

SVD Theorem: Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a rectangular matrix of rank $r \in [0, min(m,n)]$. SVD of $\boldsymbol{A}$ is a decomposition of the form:



with an orthogonal matrix $\mathbf{U} \in \mathbb{R}^{m\times m}$ with column vectors $u_i, i = 1, 2, \dots m$.

and an orthogonal matrix $\mathbf{V} \in \mathbb{R}^{n\times n}$ with column vectors $v_j, j = 1, 2,$ …n.

Also, $\boldsymbol{\Sigma}$ is an $m \times n$ matrix with $\boldsymbol{\Sigma}_{ii} = \sigma_i \geq 0$ and $\boldsymbol{\Sigma}_{ij} = 0, i \neq j$.

$\sigma_i, i = 1,2, \dots r$ of $\boldsymbol{\Sigma}$ are called the **singular values**,

u_i are called the **left-singular vectors**, and

v_j are called the **right-singular vectors.**

Singular values are ordered, i.e., $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$

The singular value matrix $\boldsymbol{\Sigma}$ is unique.

Observe that the $\boldsymbol{\Sigma} \in \mathbb{R}^{\boldsymbol{m \times n}}$ is of the same size as $\boldsymbol{A}$.

This means that $\boldsymbol{\Sigma}$ has a diagonal submatrix that contains the singular values and needs additional zero padding.

If $m > n$, then, $\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$

If $m < n$, then, $\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots & & \vdots \\ 0 & 0 & \sigma_m & 0 & \cdots & 0 \end{bmatrix}$

Remark. The SVD exists for any matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$.

Geometric Intuitions for the SVD

SVD gives geometric intuitions to describe a transformation matrix A.

SVD of a matrix can be interpreted as a decomposition of a corresponding linear mapping into three operations

SVD can be described as sequential linear transformations performed on the bases.

Geometric Interpretation of SVD

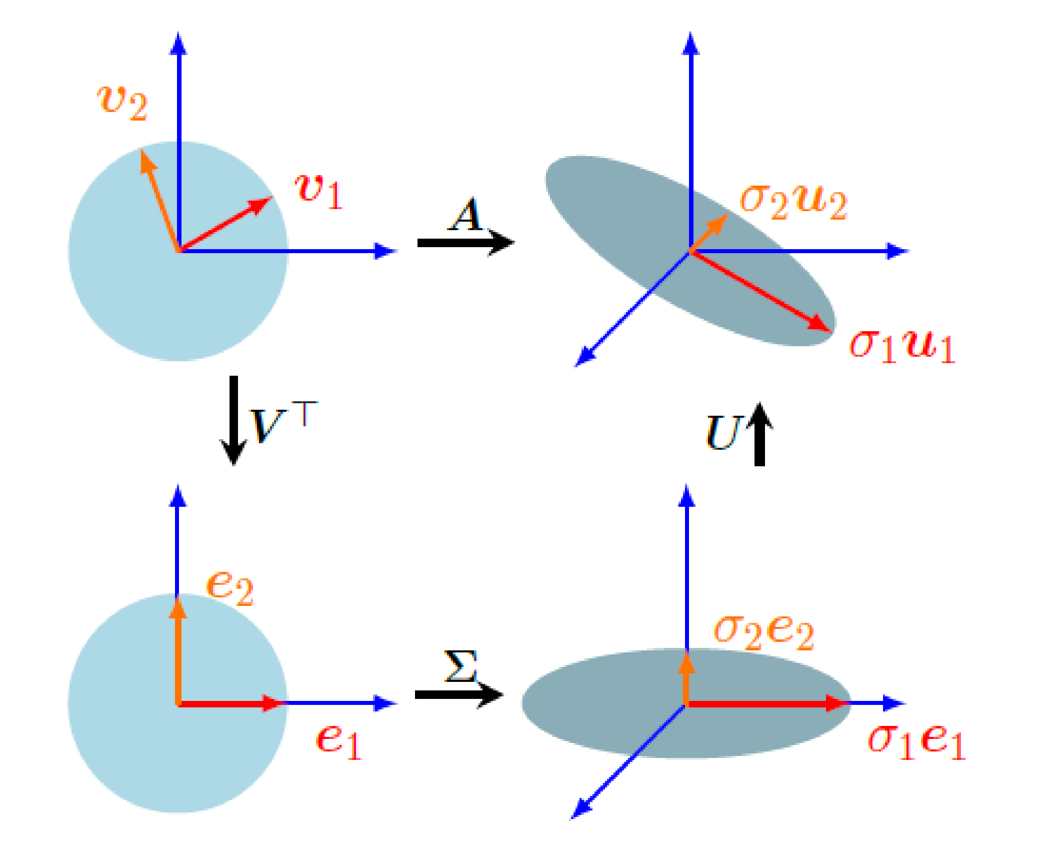
SVD on $\boldsymbol{A} \in \mathbb{R}^{3\times2}$ as sequential transformations.

Top-left to bottom-left: V^T performs a basis change in $\mathbb{R}^2$.

Bottom-left to bottom-right: Σ scales and maps from $\mathbb{R}^2$ to $\mathbb{R}^3$.

The ellipse in the bottom-right lives in $\mathbb{R}^3$, where the third dimension is orthogonal to surface of elliptical disk.

Bottom-right to top-right: U performs a basis change within $\mathbb{R}^3$.



Given a transformation matrix of a linear mapping $\Phi: \mathbb{R}^n \longrightarrow \mathbb{R}^m$ w.r.t. standard bases $\boldsymbol{B}$ and $\boldsymbol{C}$ of $\mathbb{R}^n$ and $\mathbb{R}^m$, resp.

Also, let there be second basis $\widetilde{\boldsymbol{B}}$ of $\mathbb{R}^n$ and $\widetilde{\boldsymbol{C}}$ of $\mathbb{R}^m$. Then:

1. The matrix $\boldsymbol{V}$ performs a basis change in the domain $\mathbb{R}^n$ from $\widetilde{\boldsymbol{B}}$ (red and orange vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ in the top-left of Fig) to the standard basis $\boldsymbol{B}$. Also, $V^T = V^{-1}$ performs a basis change from $\boldsymbol{B}$ to $\widetilde{\boldsymbol{B}}$. The red and orange vectors are now aligned with the canonical basis in the bottom-left of Fig.

2. Having changed the coordinate system to $\widetilde{\boldsymbol{B}}$, $\boldsymbol{\Sigma}$ scales the new coordinates by the singular values σ_i (and adds or deletes dimensions), i.e., $\boldsymbol{\Sigma}$ is transformation matrix of Φ w.r.t. $\widetilde{\boldsymbol{B}}$ and $\widetilde{\boldsymbol{C}}$, resp. by the red and orange vectors being stretched and lying in the $e_1 - e_2$ plane, which is now embedded in a third dimension in the bottom-right of Fig.

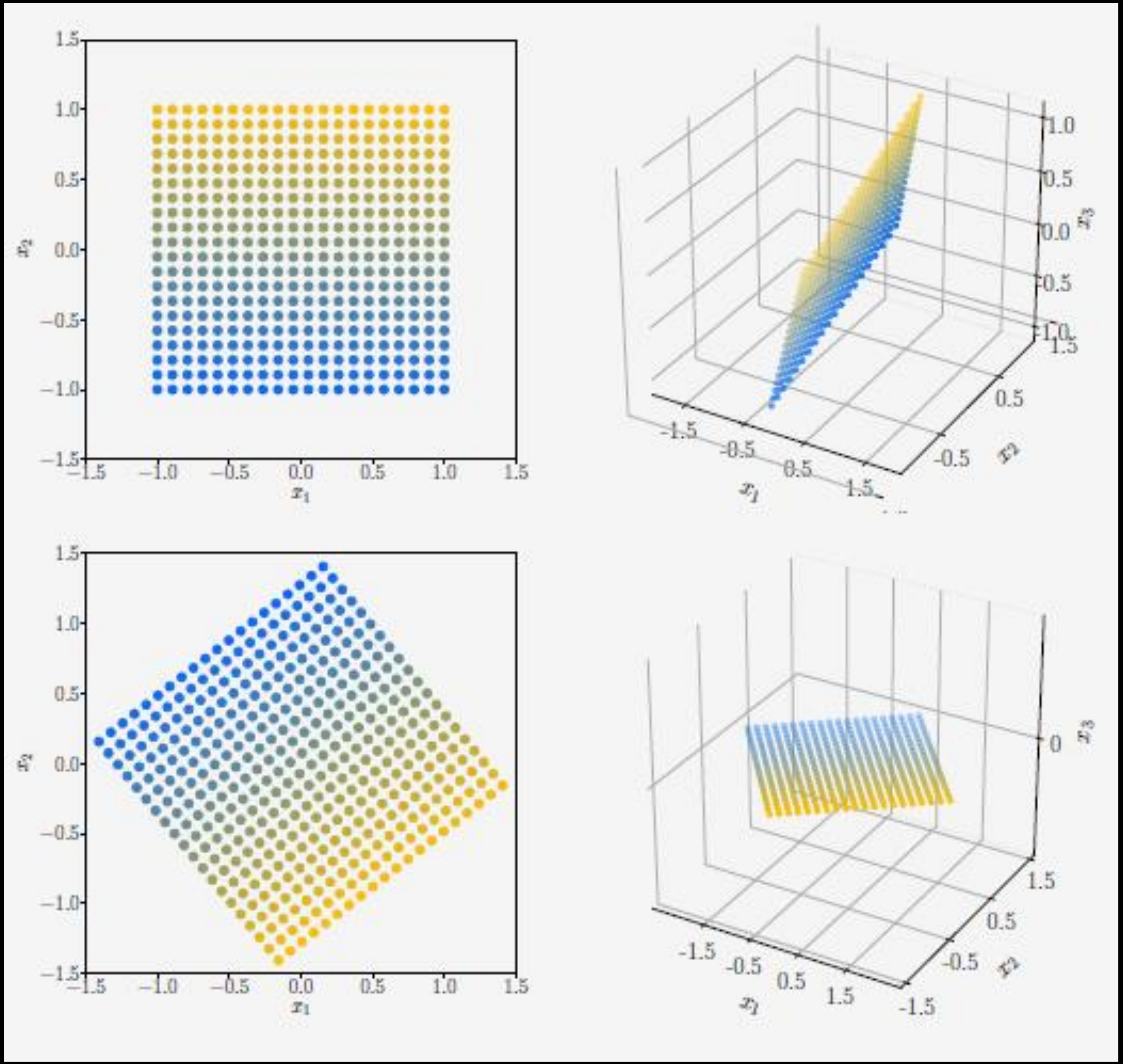
3. $\boldsymbol{U}$ performs a basis change in the codomain $\mathbb{R}^m$ from $\widehat{\boldsymbol{C}}$ into the canonical basis of $\mathbb{R}^m$, rep. by a rotation of the red and orange vectors out of the $e_1 - e_2$ plane. This is shown in the top-right of Fig.

The SVD expresses a change of basis in both the domain and codomain. This is in contrast with the eigendecomposition that operates within the same vector space, where the same basis change is applied and then undone. What makes the SVD special is that these two different bases are simultaneously linked by the singular value matrix $\boldsymbol{\Sigma}$.

Ex: Consider a mapping of a square grid of vectors $\boldsymbol{X} \in \mathbb{R}^2$ that fit in a box of size 2×2 centered at the origin. Using the standard basis, we map these vectors using:

$$\boldsymbol{A} = \begin{bmatrix} 1 & -0.8 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \boldsymbol{U\Sigma V^T}$$

$$= \begin{bmatrix} -0.79 & 0 & -0.62 \\ 0.38 & -0.78 & -0.49 \\ -0.48 & -0.62 & 0.62 \end{bmatrix} \begin{bmatrix} 1.62 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -0.78 & 0.62 \\ -0.62 & -0.78 \end{bmatrix}$$


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
x_2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
x_1
1.0
0.5
0.0
0.5
1.0
x_3
1.5
0.5
−0.5
x_2
−1.5
−0.5
0.5
1.5
x_1
0
x_3
-1.5

Start with a set of vectors $\boldsymbol{X}$ (colored dots) arranged in a grid.

Apply $\boldsymbol{V}^T \in \mathbb{R}^{2\times 2}$, which rotates $\boldsymbol{X}$.

The rotated vectors are shown in the bottom-left panel of Fig.

Map these vectors using the singular value matrix $\boldsymbol{\Sigma}$ to the codomain $\mathbb{R}^3$ (see the bottom-right panel in Fig.).

Note that all vectors lie in the $x_1 - x_2$ plane. The third coordinate is always 0. The vectors in the $x_1 - x_2$ plane have been stretched by the singular values.

The direct mapping of the vectors $\boldsymbol{X}$ by $\boldsymbol{A}$ to the codomain $\mathbb{R}^3$ equals the transformation of $\boldsymbol{X}$ by $\boldsymbol{U\Sigma V^T}$, where $\boldsymbol{U}$ performs a rotation within the codomain $\mathbb{R}^3$ so that the mapped vectors are no longer restricted to the $x_1 - x_2$ plane; they still are on a plane as shown in the top-right panel of Fig.

Construction of the SVD

Note:

Compare the eigendecomposition of an SPD matrix

$$\boldsymbol{S} = \boldsymbol{S}^T = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T$$

with the corresponding SVD

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

If we set $\boldsymbol{U} = \boldsymbol{P} = \boldsymbol{V}$ and $\boldsymbol{D} = \boldsymbol{\Sigma}$, then the Singular Value Decomposition of Symmetric Positive Definite (SPD) matrices is their eigendecomposition.

How the SVD is constructed

Computing the SVD of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is equivalent to finding two sets of orthonormal bases $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \dots \boldsymbol{u}_m)$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \dots\ \boldsymbol{v}_n)$ of the codomain $\mathbb{R}^m$ and the domain $\mathbb{R}^n$, resp. From these ordered bases, matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ will be constructed.

1. Construct the orthonormal set of right singular vectors $(\boldsymbol{v}_1, \boldsymbol{v}_2, \dots\ \boldsymbol{v}_n) \in \mathbb{R}^n$.
2. Construct the orthonormal set of left-singular vectors $(\boldsymbol{u}_1, \boldsymbol{u}_2, \dots \boldsymbol{u}_m) \in \mathbb{R}^m$.
3. Link the two and require that orthogonality of $\boldsymbol{v}_i$ is preserved under the transformation of $\boldsymbol{A}$. (This is important because the images $\boldsymbol{A}\boldsymbol{v}_i$ form a set of orthogonal vectors.)
4. Normalize these images by scalar factors, which will be the singular values.

1. Construct right-singular vectors.

According to spectral theorem the eigenvectors of a symmetric matrix form an ONB, and it can be diagonalized. We can always construct a symmetric, positive semidefinite matrix $\boldsymbol{A}^T\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{n}}$ from any rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{n}}$. (slide 29, lecture 7)

Thus, we can always diagonalize $\boldsymbol{A}^T\boldsymbol{A}$ and obtain:

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T = \boldsymbol{P}\begin{bmatrix} \boldsymbol{\lambda_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{\lambda_n} \end{bmatrix}\boldsymbol{P}^T$$

where $\boldsymbol{P}$ is an orthogonal matrix, which is composed of the orthonormal eigenbasis. The $\lambda_i \geq 0$ are the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$.

So:

$$\boldsymbol{A^T A = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^T U^T U\Sigma V^T}$$

where U and V are orthogonal matrices.

$$\boldsymbol{= V\Sigma^T \Sigma V^T = V\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n^2 \end{bmatrix} V^T}$$

Comparing, $\boldsymbol{V^T = P^T}$ and $\boldsymbol{\sigma_i^2 = \lambda_i}$

Thus, the eigenvectors of $\boldsymbol{A^T A}$ that compose $\boldsymbol{P}$ are the right-singular vectors $\boldsymbol{V}$ of $\boldsymbol{A}$.

The eigenvalues of $\boldsymbol{A^T A}$ are the squared singular values of $\boldsymbol{\Sigma}$.

2. Construct the left-singular vectors $\boldsymbol{U}$

Computing the SVD of the symmetric matrix $\boldsymbol{A^T A} \in \mathbb{R}^{\boldsymbol{m\times m}}$, which is: $\boldsymbol{AA^T = (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma V^T V \Sigma^T U^T}$

$$\boldsymbol{= U\Sigma\Sigma^T U^T}$$

$$\boldsymbol{= U\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_m^2 \end{bmatrix} U^T}$$

Using spectral theorem, $\boldsymbol{AA^T = SDS^T}$ can be diagonalized and we can find an ONB of eigenvectors of $\boldsymbol{AA^T}$ which are collected in $\boldsymbol{S}$.

The orthonormal eigenvectors of $\boldsymbol{AA^T}$ are the left-singular vectors $\boldsymbol{U}$ and form an orthonormal basis in the codomain of the SVD.

3. Structure of the matrix $\boldsymbol{\Sigma}$

Since $\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$ have the same nonzero eigenvalues, the nonzero entries of the $\boldsymbol{\Sigma}$ matrices in the SVD for both cases have to be the same.

4. Connect the orthonormal set of right-singular vectors in $\boldsymbol{V}$ to the orthonormal vectors in $\boldsymbol{U}$.

The images of $\boldsymbol{v}_i$ under $\boldsymbol{A}$ have to be orthogonal as well. We can show this, we require that the inner product between $\boldsymbol{A}\boldsymbol{v}_i$ and $\boldsymbol{A}\boldsymbol{v}_j$ must be 0 for $i \neq j$.

For any two orthogonal eigenvectors $v_i, v_j, i \neq j$, it holds

$$(Av_i)^T(Av_j) = v_i^T(A^TA)v_j = v_i^T(\lambda_j v_j) = \lambda_j v_i^T v_j = 0$$

For the case $m \geq r$, it holds that $\{\boldsymbol{Av}_1, \ldots \boldsymbol{Av}_r\}$ is a basis of an r −dimensional subspace of $\mathbb{R}^m$.

We need left-singular vectors that are orthonormal.

We normalize the images of the right-singular vectors $\boldsymbol{Av}_i$ and obtain :

$$u_i = \frac{\boldsymbol{Av}_i}{\|\boldsymbol{Av}_i\|} = \frac{1}{\sqrt{\lambda_i}}\boldsymbol{Av}_i = \frac{1}{\sigma_i}\boldsymbol{Av}_i$$

Thus, the eigenvalues of $\boldsymbol{AA}^T$ are such that $\sigma_i^2 = \lambda_i$.

Therefore, the eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$, which we know are the right singular vectors $\boldsymbol{v}_i$, and their normalized images under $\boldsymbol{A}$, the left-singular vectors $\boldsymbol{u}_i$, form two self-consistent ONBs that are connected through the singular value matrix $\boldsymbol{\Sigma}$.

Rearranging we get the **<u>singular value equation</u>**

$$\boldsymbol{A}\boldsymbol{v}_{\boldsymbol{i}} = \sigma_i \boldsymbol{u}_i, i = 1, 2, \dots r.$$

This equation closely resembles the eigenvalue equation $(\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x})$ but the vectors on the left- and the right-hand sides are not the same.

For $n < m$, the singular value equation holds only for $i \leq n$, but singular value equation says nothing about the u_i for $i > n$. However, we know by construction that they are orthonormal.

Conversely, for $m < n$, the singular value equation holds only for $i \leq m$. For $i > m$, we have $\boldsymbol{A}\boldsymbol{v}_i = 0$ and we still know that the $\boldsymbol{v_i}$ form an orthonormal set. This means that the SVD also supplies an orthonormal basis of the kernel

(null space) of $\boldsymbol{A}$, the set of vectors $\boldsymbol{x}$ with $\boldsymbol{Ax} = \boldsymbol{0}$.

Concatenating the $\boldsymbol{v_i}$ as the columns of V and the $\boldsymbol{u}_i$ as the columns of U yields $\boldsymbol{AV} = \boldsymbol{U\Sigma}$, where $\boldsymbol{\Sigma}$ has the same dimensions as $\boldsymbol{A}$ and a diagonal structure for rows $1, 2, \ldots r$.

Hence, right-multiplying with $\boldsymbol{V}^T$ yields $\mathbf{A} = \boldsymbol{U\Sigma V}^T$, which is the SVD of $\boldsymbol{A}$.

Ex: Compute the SVD of $\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}$

We need to compute the Right-singular vectors v_j , the singular values σ_k, and the left singular vectors u_i.

Step 1: Right-singular vectors as the eigenbasis of $\boldsymbol{A}^T\boldsymbol{A}$.

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 5 & -2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Eigen values of $\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} 5 & -2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$

$$\left| det \begin{bmatrix} 5 & -2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right| = 0$$

$$\Leftrightarrow \left| det \begin{bmatrix} 5-\lambda & -2 & 1 \\ -2 & 1-\lambda & 0 \\ 1 & 0 & 1-\lambda \end{bmatrix} \right| = 0$$

$$\Leftrightarrow (5-\lambda)(1-\lambda)^2 - 2(-2)(1-\lambda) + (-2)(1-\lambda) = 0$$

Eigen values are: $\lambda = 6, 1\ and\ 0$

Find eigen vectors w.r.t. each of the Eigen values:

For $\lambda = 6$: $\begin{bmatrix} 5 & -2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 6 \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

Solving gives $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ {}^{-2}/_{5} \\ {}^{1}/_{5} \end{bmatrix}$

Normalizing it , gives $\begin{bmatrix} {}^{1}/_{\sqrt{30}} \\ {}^{-2}/_{\sqrt{30}} \\ {}^{1}/_{\sqrt{30}} \end{bmatrix}$

For λ =1, corresponding eigen vector is: $\begin{bmatrix} 0 \\ {}^{1}/_{\sqrt{5}} \\ {}^{2}/_{\sqrt{5}} \end{bmatrix}$

For $\lambda = 0$, corresponding eigen vector is: $\begin{bmatrix} {}^{-1}/_{\sqrt{6}} \\ {}^{-2}/_{\sqrt{6}} \\ {}^{1}/_{\sqrt{6}} \end{bmatrix}$

Next, compute the singular values and right-singular vectors $\boldsymbol{v}_j$ through the eigenvalue decomposition of $\boldsymbol{A}^T\boldsymbol{A}$, which is:

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} 5/\sqrt{30} & 0 & -1/\sqrt{6} \\ -2/\sqrt{30} & 1/\sqrt{5} & -2/\sqrt{6} \\ 1/\sqrt{30} & 2/\sqrt{5} & 1/\sqrt{6} \end{bmatrix} \begin{bmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 5/\sqrt{30} & -2/\sqrt{30} & 1/\sqrt{30} \\ 0 & 1/\sqrt{5} & 2/\sqrt{5} \\ -1/\sqrt{6} & -2/\sqrt{6} & 1/\sqrt{6} \end{bmatrix}$$

$$= \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T$$

And the right-singular vectors as the columns of $\boldsymbol{P}$ so that:

$$\boldsymbol{V} = \boldsymbol{P} = \begin{bmatrix} {}^{5}/_{\sqrt{30}} & 0 & {}^{-1}/_{\sqrt{6}} \\ {}^{-2}/_{\sqrt{30}} & {}^{1}/_{\sqrt{5}} & {}^{-2}/_{\sqrt{6}} \\ {}^{1}/_{\sqrt{30}} & {}^{2}/_{\sqrt{5}} & {}^{1}/_{\sqrt{6}} \end{bmatrix}$$

Step 2: Singular-value matrix.

As the singular values σ_i are the square roots of the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$, we obtain them straight from $\boldsymbol{D}$. Since $rk(\boldsymbol{A}) = 2$, there are only two nonzero singular values: $\sigma_1 = \sqrt{6}$ and $\sigma_2 = 1$. The singular value matrix must be the same size as $\boldsymbol{A}$, and we get:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Step 3: Left-singular vectors as the normalized image of the rightsingular vectors.

We find the left-singular vectors by computing the image of the rightsingular vectors under $\boldsymbol{A}$ and normalizing them by dividing them by their corresponding singular value. We obtain

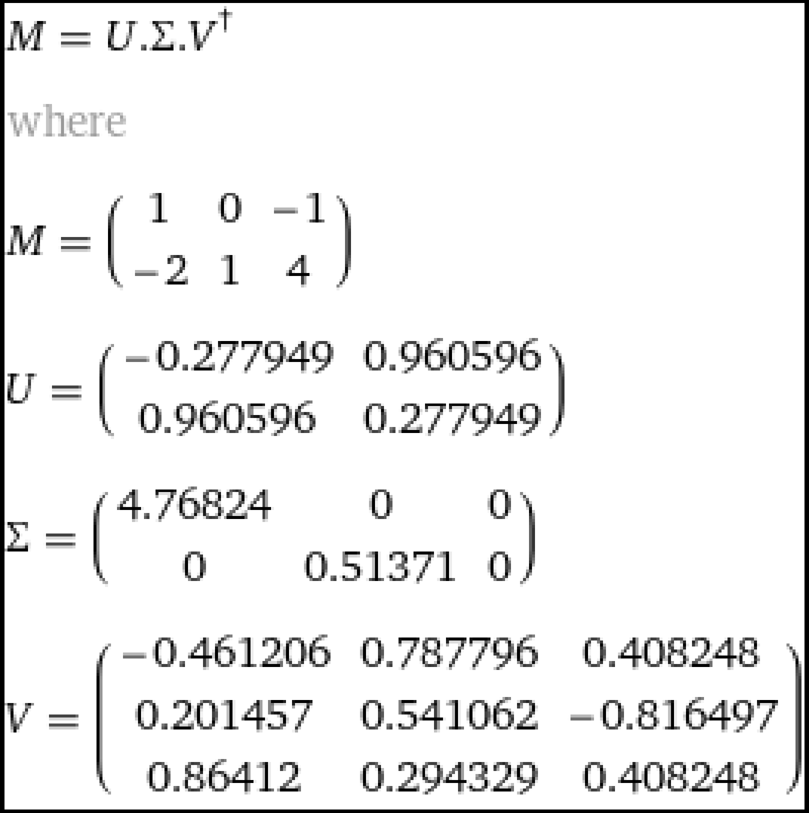
$$u_1 = \frac{1}{\sigma_1} A v_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 5/\sqrt{30} \\ -2/\sqrt{30} \\ 1/\sqrt{30} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{bmatrix}$$

$$u_2 = \frac{1}{\sigma_2} A v_2 = \frac{1}{1} \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}$$

And $\boldsymbol{U} = [u_1, u_2] = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}$

Ex: Compute the SVD of $\boldsymbol{A} = \begin{bmatrix} 1 & 0 & -1 \\ -2 & 1 & 4 \end{bmatrix}$

Ans:



$$M = U.\Sigma.V^{\dagger}$$

where

$$M = \begin{pmatrix} 1 & 0 & -1 \\ -2 & 1 & 4 \end{pmatrix}$$

$$U = \begin{pmatrix} -0.277949 & 0.960596 \\ 0.960596 & 0.277949 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 4.76824 & 0 & 0 \\ 0 & 0.51371 & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} -0.461206 & 0.787796 & 0.408248 \\ 0.201457 & 0.541062 & -0.816497 \\ 0.86412 & 0.294329 & 0.408248 \end{pmatrix}$$

Ex: **Movie Ratings and Consumers**

Viewers: Ali, Beatrix, Chandra

Movies: Star Wars, Blade Runner, Amelie, Delicatessen

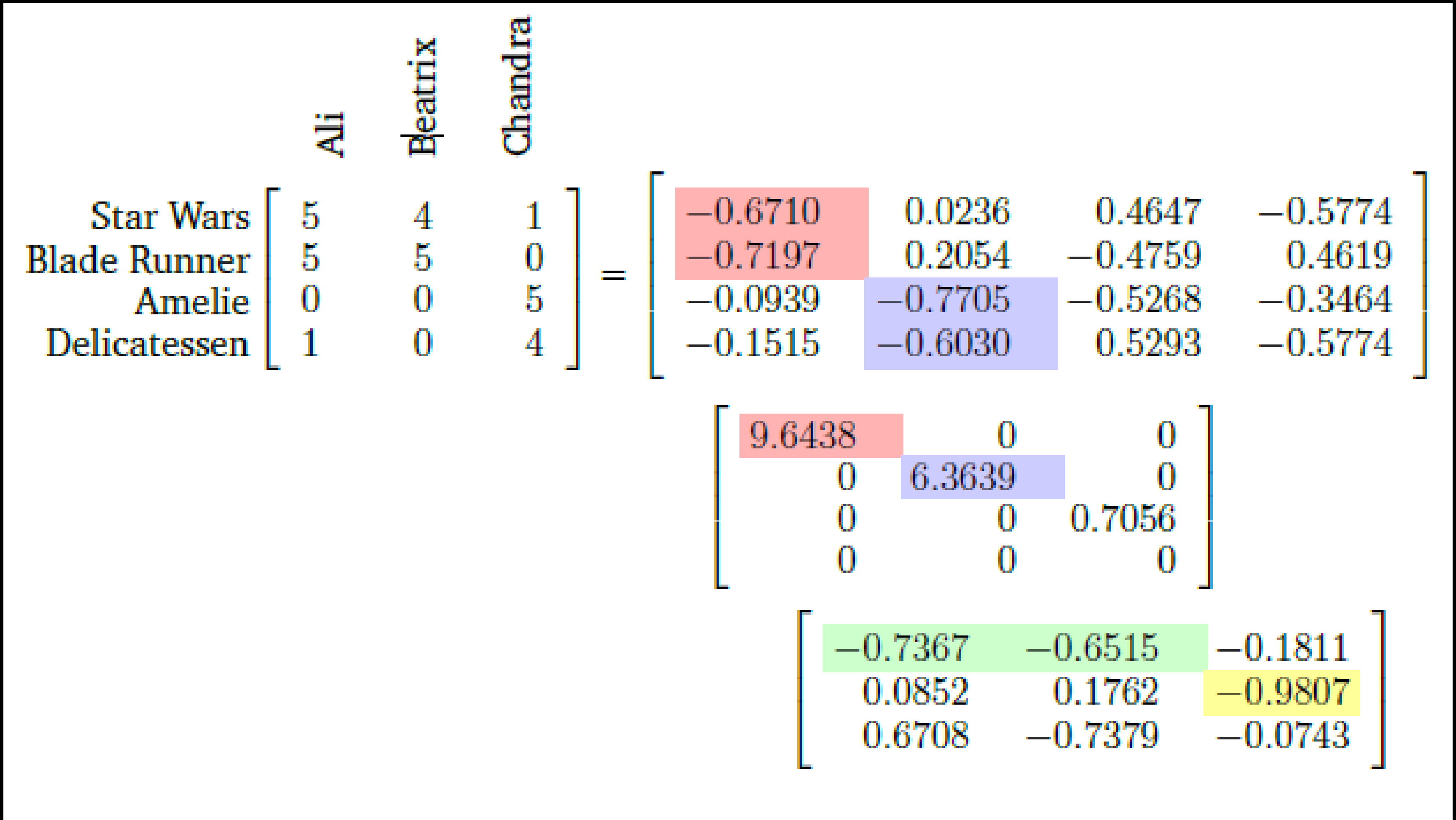
Ratings are between 0 (worst) and 5 (best)

Data is in a matrix $\boldsymbol{A} \in \mathbb{R}^{\mathbf{4\times3}}$.

Assumptions: Any viewer's specific movie preferences can be expressed as a linear combination of the v_j.

Any movie's like-ability can be expressed as a linear combination of the u_i.

Factoring $\boldsymbol{A}$ using the SVD provides the relationships of how people rate movies, especially if there is a structure linking which people like which movies.



$$\begin{array}{r|ccc|}
 & \text{Ali} & \text{Beatrix} & \text{Chandra} \\
\text{Star Wars} & 5 & 4 & 1 \\
\text{Blade Runner} & 5 & 5 & 0 \\
\text{Amelie} & 0 & 0 & 5 \\
\text{Delicatessen} & 1 & 0 & 4
\end{array} = \begin{bmatrix} -0.6710 & 0.0236 & 0.4647 & -0.5774 \\ -0.7197 & 0.2054 & -0.4759 & 0.4619 \\ -0.0939 & -0.7705 & -0.5268 & -0.3464 \\ -0.1515 & -0.6030 & 0.5293 & -0.5774 \end{bmatrix}$$

$$\begin{bmatrix} 9.6438 & 0 & 0 \\ 0 & 6.3639 & 0 \\ 0 & 0 & 0.7056 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} -0.7367 & -0.6515 & -0.1811 \\ 0.0852 & 0.1762 & -0.9807 \\ 0.6708 & -0.7379 & -0.0743 \end{bmatrix}$$

Interpretation:

The first left-singular vector u_1 has large absolute values for the two science fiction movies and a large first singular value (red shading). Thus, this groups a type of users with a specific set of movies (science fiction theme).

Similarly, the first right-singular v_1 shows large absolute values for Ali and Beatrix, who give high ratings to science fiction movies (green shading). Thus, v_1 reflects the notion of a science fiction lover.

Similarly, u_2, seems to capture a French art house film theme, and v_2 indicates that Chandra is close to an idealized lover of such movies. An idealized science fiction lover is a purist and only loves science fiction movies, so a science fiction lover v_1 gives a rating of zero to everything but science fiction themed—this logic is implied by the diagonal substructure for the singular value matrix $\boldsymbol{\Sigma}$.

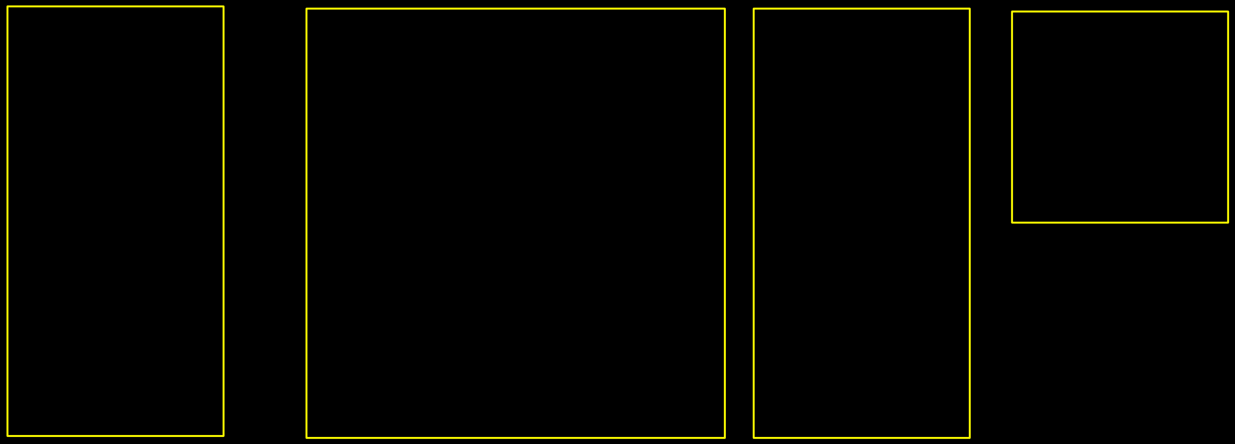
A specific movie is therefore represented by how it decomposes (linearly) into its stereotypical movies. Likewise, a person would be represented by how they decompose (via linear combination) into movie themes.

SVD terminology and conventions

Full SVD

Two square left- and right-singular vector matrices, but a non-square singular value matrix

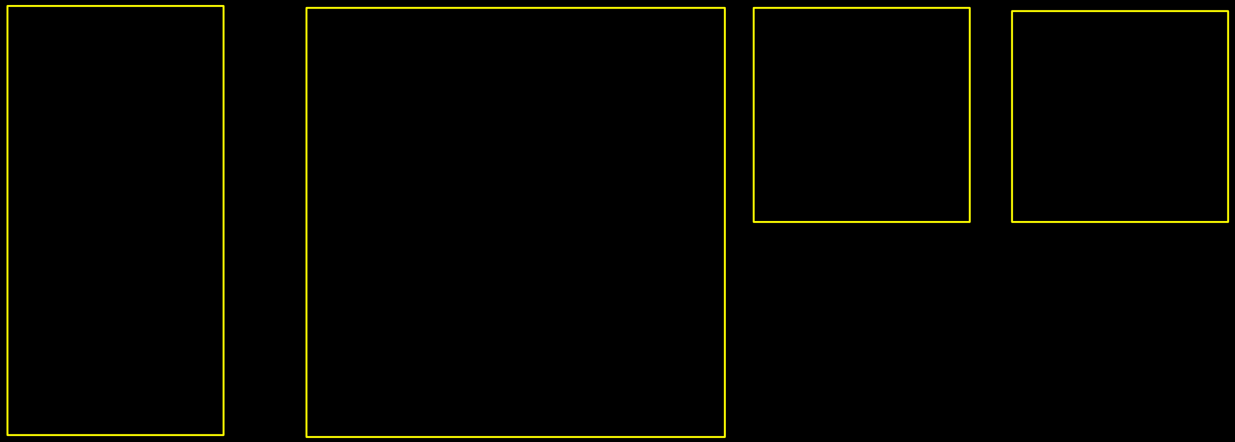
$$(A)_{m\times n} = (\boldsymbol{U})_{m\times m} \quad (\boldsymbol{\Sigma})_{m\times n} \quad \boldsymbol{V}^T_{n\times n}$$



Reduced SVD

For $A \in \mathbb{R}^{m\times n}$ and $m \geq n$

$$(A)_{m\times n} = (\boldsymbol{U})_{m\times n} \quad (\boldsymbol{\Sigma})_{n\times n} \quad \boldsymbol{V}^T_{n\times n}$$



Truncated SVD

Matrix approximation technique using the SVD

SVD of a rank$-r$ matrix A

$$(A)_{m\times n} = (\boldsymbol{U})_{m\times r}(\boldsymbol{\Sigma})_{r\times r}\boldsymbol{V}_{r\times n}^{T}$$

$(\boldsymbol{\Sigma})_{r\times r}$ is a diagonal matrix having only nonzero entries along the diagonal. (just like eigenvalue decomposition).

When $m < n$, the SVD decomposition will yield $\boldsymbol{\Sigma}$ with more zero columns than rows and, so, the singular values $\sigma_m, \sigma_{m+1}, \ldots\ \sigma_n$ are 0.

Some interesting facts about SVD

Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{A} = \boldsymbol{U\Sigma V}^T$, where $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ be orthogonal matrices. Then:

1. $rank(\boldsymbol{A}) = rank(\boldsymbol{\Sigma}) = r$
2. Column space of $\boldsymbol{A}$ is spanned by first r columns of $\boldsymbol{U}$.
3. Null space of $\boldsymbol{A}$ is spanned by last $n - r$ columns of $\boldsymbol{V}$.
4. Row space of $\boldsymbol{A}$ is spanned by first r columns of $\boldsymbol{V}$.
5. Null space of $\boldsymbol{A}^T$ is spanned by last $m - r$ columns of $\boldsymbol{U}$.

Applications of SVD

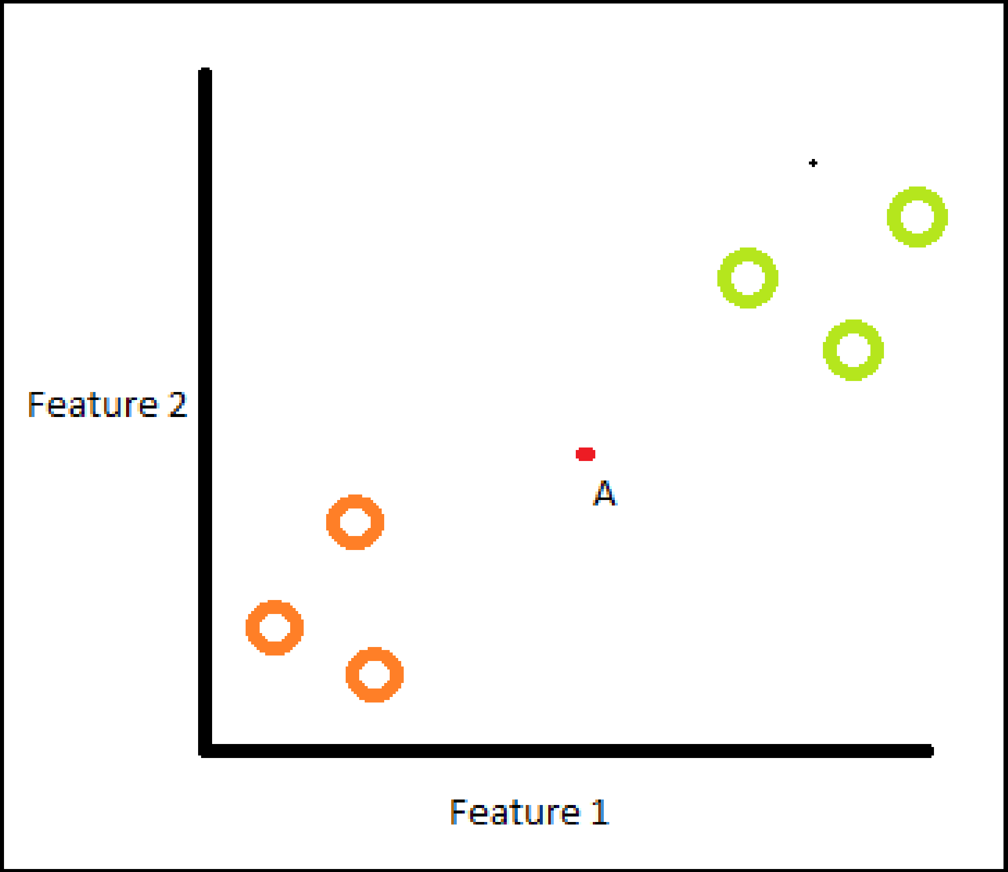
- least-squares problems in curve fitting
- Solution of systems of linear equations
- Dimensionality reduction
- Data compression
- Clustering

Principle Component Analysis

PCA is used for dimensionality reduction

1. Definitions and the idea behind PCA
2. Example of how to reduce a 2-dim data to 1-dim data.
3. Example of how to reduce a 4-dim data to a 2-dim data.
4. Generalized PCA steps
5. Example using python

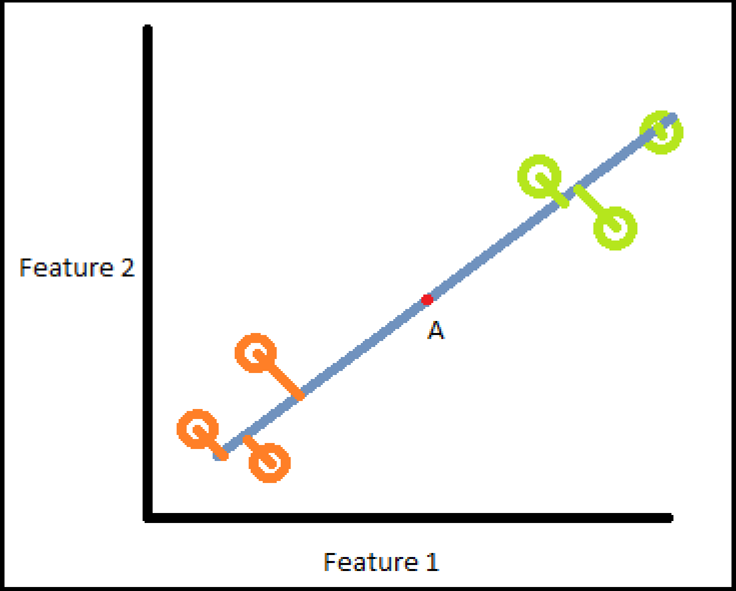
Ex: Given a 2-dimesnsional data having two features- feature 1 and feature 2.



PCA finds the best fit line for these data points which **minimizes** the distance between the data points and their projections on the best fit line.

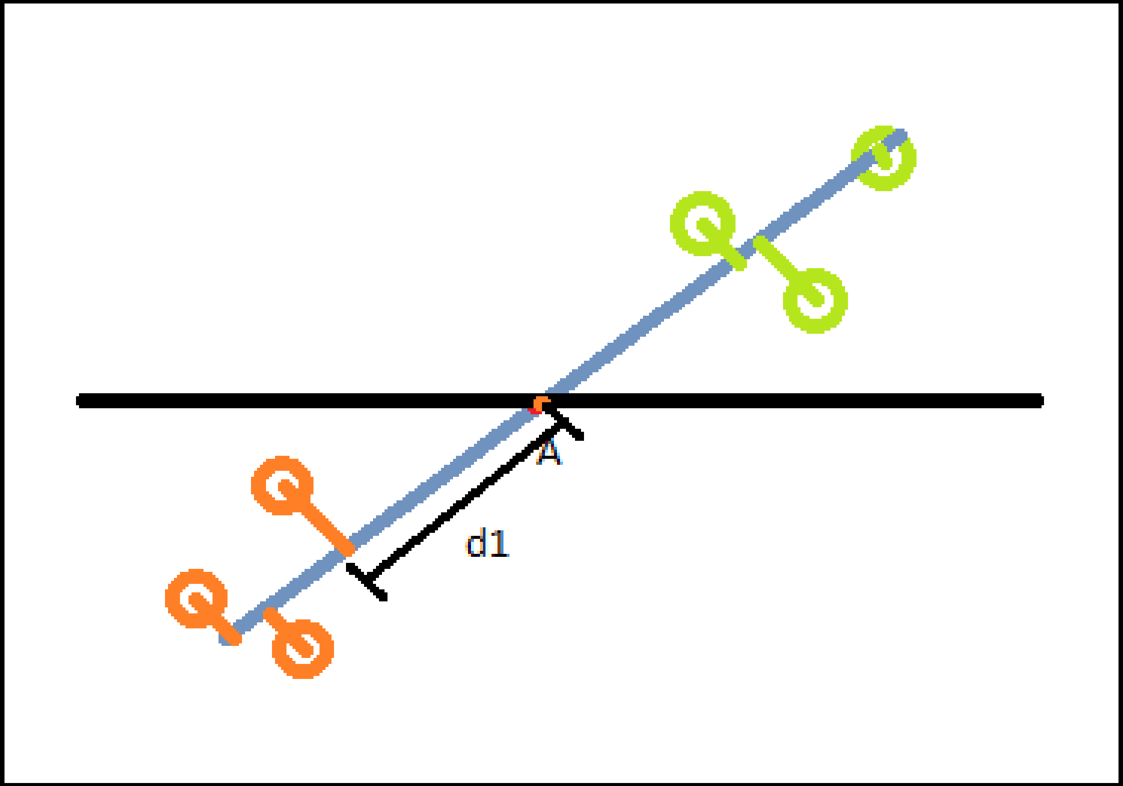
Consider the average of the data points of feature 1 and feature 2. It will be around **A**.

PCA can also **maximize** the distance of the projected points on the best fit line from the point A.



Shift the line so that the point **A** coincides with the origin, which will make it easier to visualize.

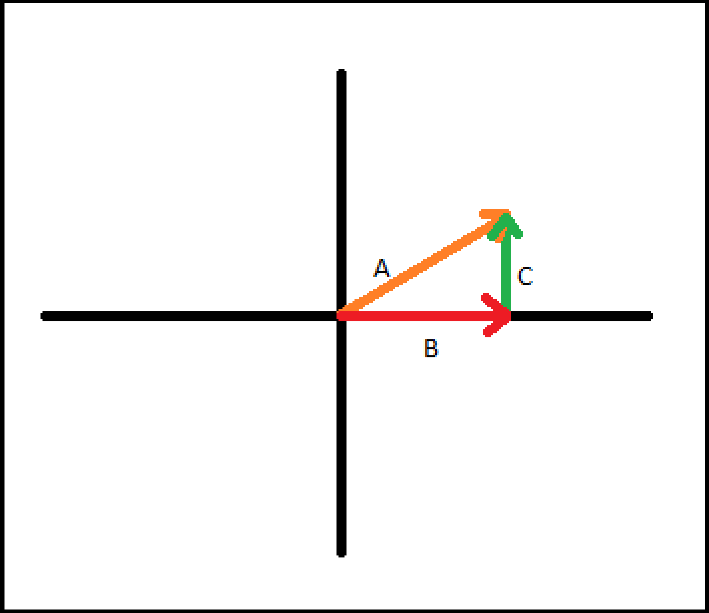
The distance d1 is the distance of the point 1 with respect to the origin.



Similarly, d2,d3,d4,d5,d6 will be the respective distances of the projected points from the origin. The best fit line will have the maximum Sum of Squares of Distances.

Let the slope of line be 0.25. That means the line consists of 4 parts of feature 1 and 1 part of feature 2, where B=4 & C=1.

Using Pythagoras Theorem, A=4.12.



PCA scales these values so that the vector A is unit length long.

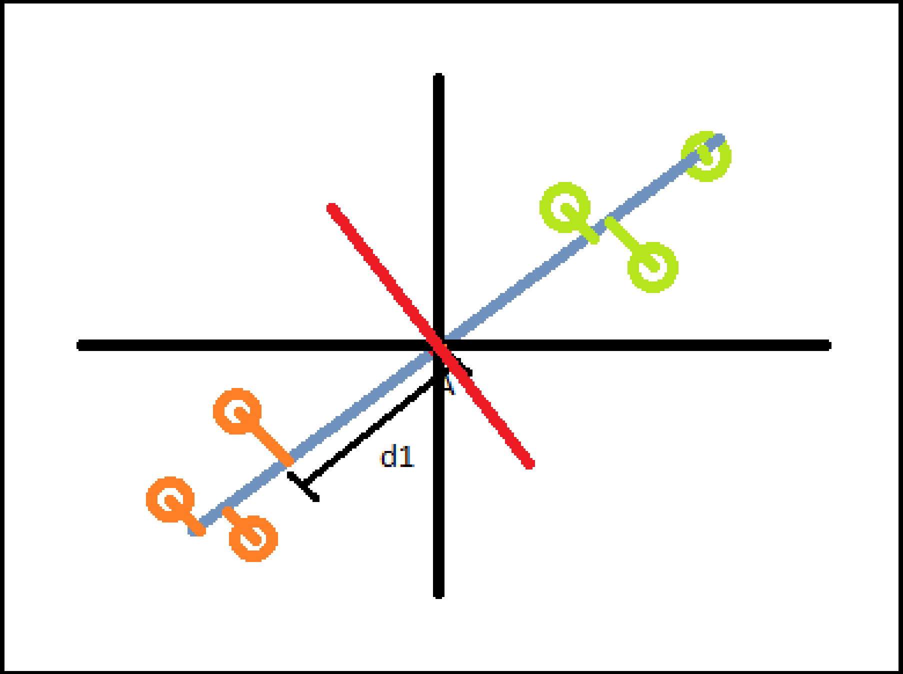
Hence A=1, B=4/4.12 = 0.97 & C=1/4.12 = 0.242. This unit vector A is the **eigenvector!**

The Sum of Squared Distances d1,d2,d3,d4,d5,d6 is the **eigenvalue.** This is the **linear combination** of feature 1 and feature 2.

This means that for PC1, feature 1 is almost 4 times as important than feature 2, i.e. it contains almost 4 times more spread(variation) in data than feature 2.

Now, the Principal Component 2 will be the vector orthogonal to PC1 as the principal components have 0 correlation among them. See red line.

Similarly, PC2 will have -0.242 parts of feature 1 and 0.97 parts of feature 2. Thus, for PC2, feature 2 is almost 4 times as important than feature 1. The eigenvector and the eigenvalue can be calculated similarly for PC2.



Interpretation in terms of Variance

The variance for the respective principal component, will be obtained on dividing Sum of Squared Distances for both the principal components values by n-1 (where n is the sample size).

Let variance for PC1 = 15 and that for PC2 = 3.

Hence the total variation around both PC = 18.

So PC1 accounts for 15/18 = 0.83 or 83% of the total variance in the data.

And PC2 accounts for 3/18 = 0.17 or 17% of the total variance in the data.

This is **Explained Variance Ratio**.

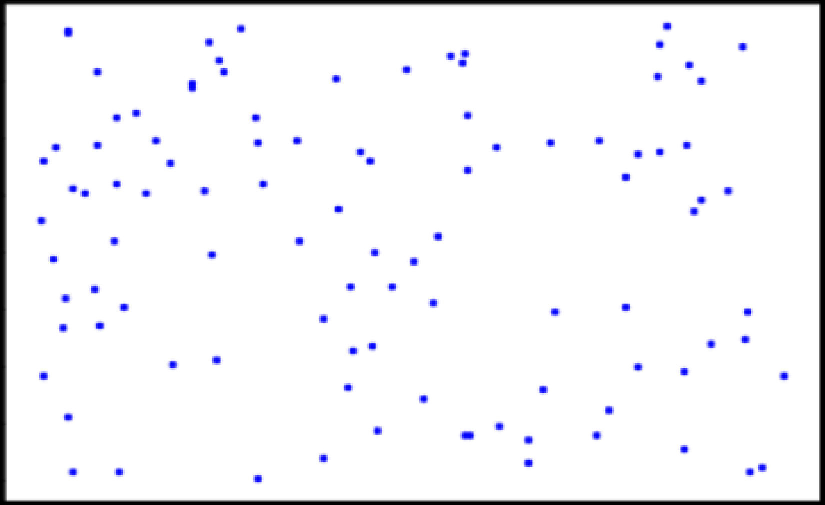
It tells how much variance in the data is explained by a particular PC. Principal components are ranked in order of their explained variance ratio. We can select top m components if the total explained variance ratio reaches a sufficient value.

Principal Component Analysis reduces the dimensionality to overcome overfitting. Your model might not need all the features to give a good performance. It might give a great training score but a very low test score. In other words, it might **overfit**. PCA is not a feature selection or a feature elimination technique.

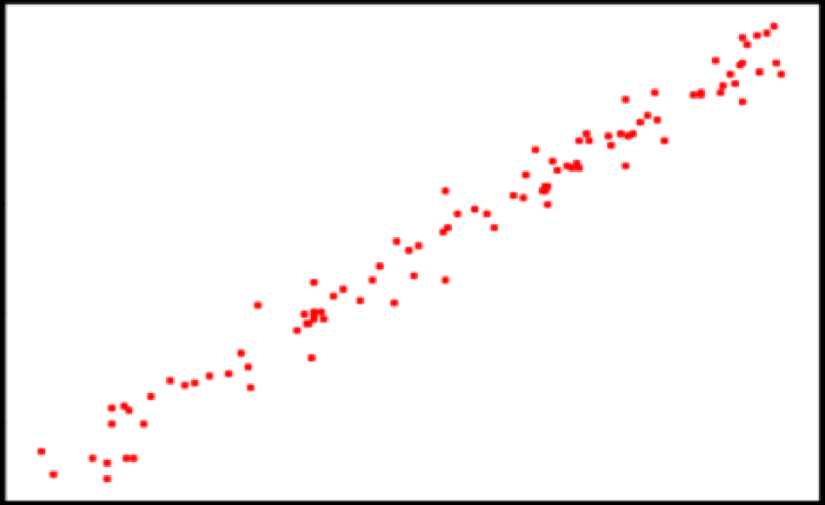
It is more of a **feature extraction** technique.

Map or transform the original features in to another feature space of smaller dimensionality. This “feature transformation approach” where the new features are constructed by applying a *linear transformation* on the original set of features is called **Principal Component Analysis.**

The use of PCA does not require knowledge of the class labels associated with each data vector. Thus, PCA is characterized as
a *linear, unsupervised* technique for dimensionality reduction.



A



B

Fig. A: No relationship or correlation in how X-Y values are varying.

Fig. B: The Y values are moving up in a linear fashion and show good correlation. Given a X-value, the Y-value can be easily guessed. The data can be represented with a good approximation lying along a line.

Thus, we can reduce the original two-dimensional data in one dimensions, thus achieving dimensionality reduction.

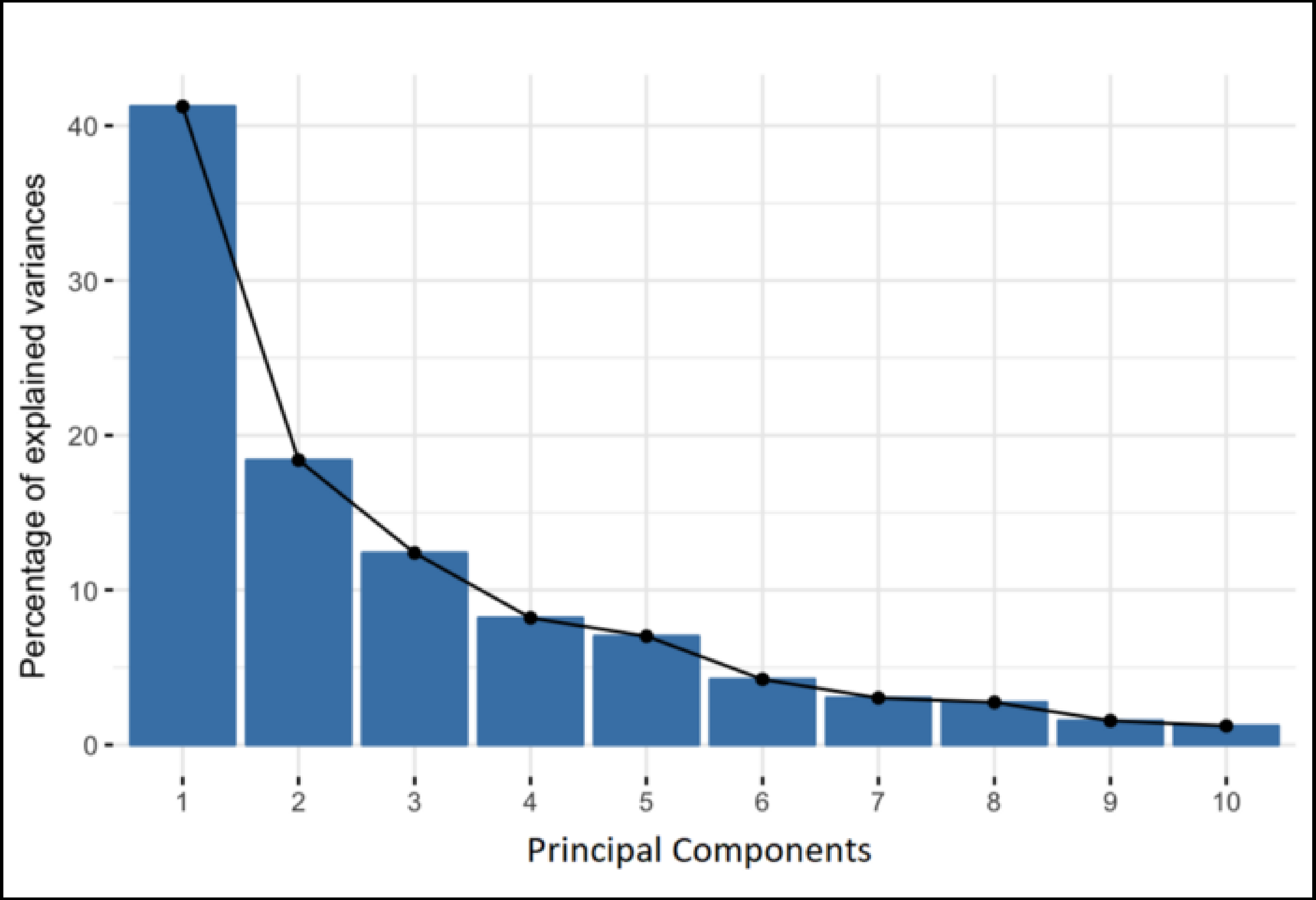
Let there be a collection of data vectors wherein each vector consists of a fixed number of attributes or features.

The number of attributes, i.e. the size of the vector, is called the **dimensionality of the feature space.**

When the dimension is large, it is often of interest to reduce the number of features, without loosing vital information in the given data set.

Principal components are new variables that are constructed as linear combinations of the initial variables, in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components.

So, a 10-dimensional data gives 10 principal components, but PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on.


Percentage of explained variances
0
10
20
30
40
1
2
3
4
5
6
7
8
9
10
Principal Components

Geometrically, principal components represent the directions of the data that explain a **maximal amount of variance**, that is, the lines that capture most information of the data. The relationship between variance and information here, is that, the larger the variance carried by a line, the larger the dispersion of the data points along it, and the larger the dispersion along a line, the more the information it has.

Thus: Principal components can be thought as new axes that provide the best angle to see and evaluate the data, so that the differences between the observations are better visible.

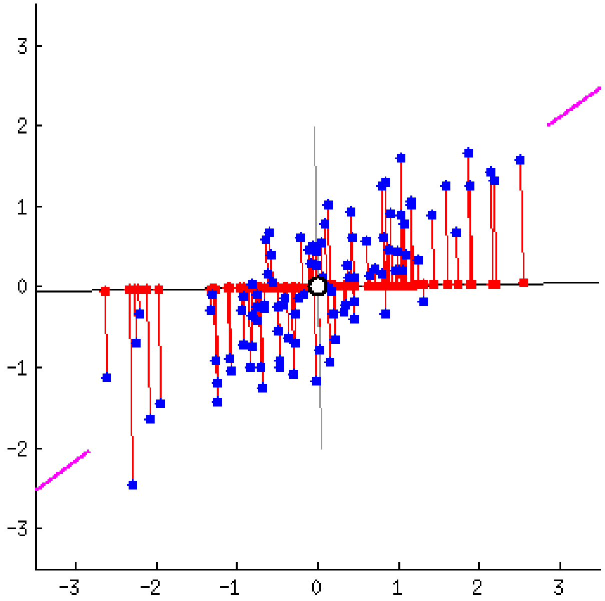
How PCA Constructs the Principal Components

Since, no. of principal components is equal to no. of variables in the data, principal components are constructed such that the first principal component accounts for the **largest possible variance** in the data set.

Example, Suppose the scatter plot of data set is given.

Then, the first principal component is approximately the line that matches the purple marks because it goes through the origin and is the line in which the projection of the points (red dots) is the most spread out.

Mathematically, it's the line that maximizes the variance (the average of the squared distances from the projected points (red dots) to the origin).


3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

The second principal component is calculated in the same way, with the condition that it is uncorrelated with (i.e., perpendicular to) the first principal component and that it accounts for the next highest variance.

Basic Definitions

Let a given sample X has values given by

$$\mathrm{X} = \{x_1, x_2, \ldots\ x_n\}$$

Def: **Mean** is defined as $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

Def: **Standard Deviation** is defined as

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Let there be two data sets

$$\mathrm{X} = \{x_1, x_2, \ldots\ x_n\} \quad \text{and} \quad \mathrm{Y} = \{y_1, y_2, \ldots\ y_n\}$$

Then **Covariance** is defined as:

$$\Sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)\,(y_i - \mu_Y)^T$$

$$\Sigma_{XY} = \begin{cases} > 0 & \quad then\ X\ and\ Y\ are\ changing\ in\ the\ same\ direction \\ 0 & \quad then\ X\ and\ Y\ are\ \ independent\ of\ each\ other \\ < 0 & \quad then\ X\ and\ Y\ are\ changing\ in\ the\ opposite\ direction \end{cases}$$

For a k-dimensional data set $\{X_1, X_2, \ldots\ X_k\}$

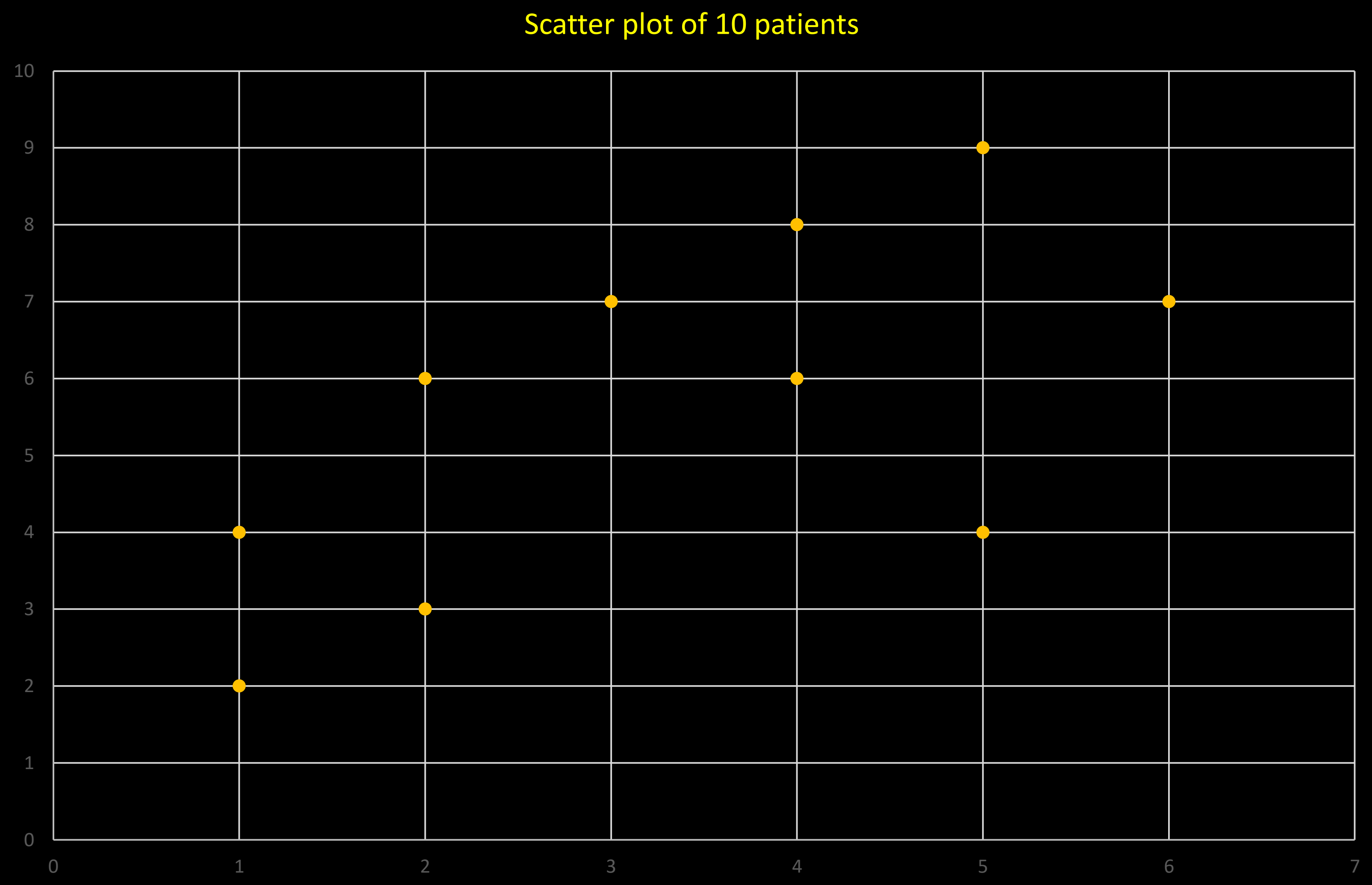
Covariance Matrix is defined as:

$$\Sigma = \begin{bmatrix} \sigma_{X_1}^2 & \Sigma_{X_1 X_2} & \cdots & \Sigma_{X_1 X_k} \\ \Sigma_{X_2 X_1} & \sigma_{X_2}^2 & \cdots & \Sigma_{X_2 X_k} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{X_k X_1} & \Sigma_{X_k X_2} & \cdots & \sigma_{X_k}^2 \end{bmatrix}$$

This is a $k \times k$ symmetric matrix having real eigen values and hence it will have orthogonal eigen vectors.

Ex 1: Let the health record of 10 patients be given on a 10-point scale with respect to BP and Sugar ailments

Patients	BP	Sugar
P1	1	2
P2	4	8
P3	2	6
P4	6	7
P5	5	9
P6	4	6
P7	2	3
P8	3	7
P9	5	4
P10	1	4
Mean	3.3	5.6
Standard Deviation	1.76	2.2706


Scatter plot of 10 patients
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7

Adjusted data matrix by subtracting mean

Patients	BP	Sugar
P1	-2.3	-3.6
P2	0.7	2.4
P3	-1.3	0.4
P4	2.7	1.4
P5	1.7	3.4
P6	0.7	0.4
P7	-1.3	-2.6
P8	-0.3	1.4
P9	1.7	-1.6
P10	-2.3	-1.6

Covariance matrix is $=\begin{bmatrix} 2.81 & 2.32 \\ 2.32 & 4.64 \end{bmatrix}$

Eigen values and corresponding eigen vectors are:

$$\lambda_1 = 1.23108\ ;\ v_1 = \begin{bmatrix} -1.4693 \\ 1 \end{bmatrix}$$

and $\lambda_2 = 6.2189\ ;\ v_2 = \begin{bmatrix} 0.68056 \\ 1 \end{bmatrix}$

λ_2 is the largest eigen value. So, v_2 is the Principal Component 1 direction.

Transformed data in 1-dim can be obtained as:

$$y_i = [0.68056 \quad 1]\begin{bmatrix} x_1^i \\ x_2^i \end{bmatrix}$$

Patients	Transformed data
P1	$0.68056 \times (-2.3) + (-3.6) = -5.16528$
P2	2.876392
P3	-0.484728
P4	3.237512
P5	4.556952
P6	0.876392
P7	-3.484728
P8	1.195832
P9	-0.443048
P10	-3.165288

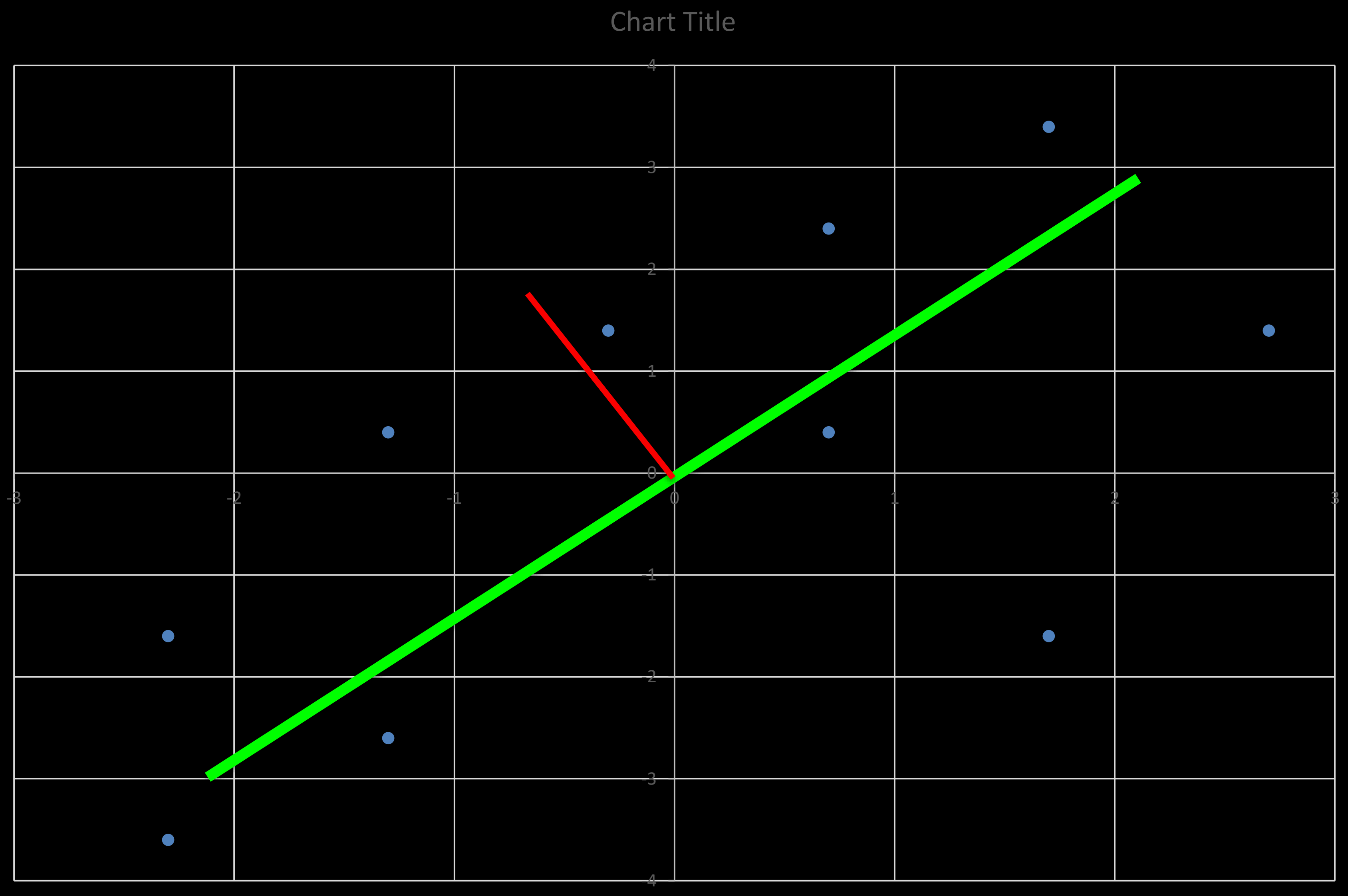


Chart Title
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3

Ex 2: Let the health record of 5 patients be given on a 10-point scale:

	BP	**Sugar**	**Obese**	**Eyesight**
Arun	4	8	4	3
Rajiv	2	6	9	3
Geeta	6	1	7	4
Ramesh	5	9	6	3
Sunder	4	6	3	4

To transform data from 4-dim to 3-dim, i.e. $\mathbb{R}^4 \rightarrow \mathbb{R}^3$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

That is, $Y_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + a_{14}X_4$

And $Y_2 = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + a_{24}X_4$

$Y_3 = a_{31}X_1 + a_{32}X_2 + a_{33}X_3 + a_{34}X_4$

To transform data from 4-dim to 2-dim, i.e. $\mathbb{R}^4 \to \mathbb{R}^2$

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

That is, $Y_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + a_{14}X_4$

And $Y_2 = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + a_{24}X_4$

Mean of given data:

$$\mu_{BP} = \frac{1}{5}(2 + 4 + 6 + 5 + 4) = 4.2$$

$$\mu_{sugar} = \frac{1}{5}(8 + 6 + 1 + 9 + 6) = 6$$

$$\mu_{obese} = \frac{1}{5}(4 + 9 + 7 + 6 + 3) = 5.8$$

$$\mu_{eyesight} = \frac{1}{5}(3 + 3 + 4 + 3 + 4) = 3.4$$

Standard Deviation of data

$$\sigma_{BP} = \sqrt{\frac{(2-4.2)^2 + \ldots + (4-4.2)^2}{5}} = \sqrt{1.76}$$

$$\sigma_{sugar} = \sqrt{7.6} = 2.7568$$

$$\sigma_{obese} = \sqrt{4.56} = 2.1354$$

$$\sigma_{eyesight}\sqrt{0.24} = 0.4898$$

In our example, covariance matrix is:

$$\Sigma = \begin{bmatrix} \sigma_{BP}^2 & \Sigma_{BP\ sugar} & \cdots & \Sigma_{BP\ eyesight} \\ \Sigma_{sugar\ BP} & \sigma_{sugar}^2 & \cdots & \Sigma_{sugar\ eyesight} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{eyesight\ BP} & \Sigma_{eyesight\ sugar} & \cdots & \sigma_{eyesight}^2 \end{bmatrix}$$

$$= \begin{bmatrix} 1.76 & -1.4 & -0.76 & 0.32 \\ -1.4 & 7.6 & -1.8 & -1 \\ -0.76 & -1.8 & 4.56 & -0.32 \\ 0.32 & -1 & -0.32 & 0.24 \end{bmatrix}$$

Def: The **Principal Components** are the eigen vectors of the covariance matrix of the given data.

In our ex: The eigen values and their eigen vectors are:

$$\lambda_1 = 0.032; v_1 = \begin{bmatrix} 0.012 \\ 0.167 \\ 0.139 \\ 1 \end{bmatrix}; \quad \lambda_2 = 1.1041; \ v_2 = \begin{bmatrix} -9.047 \\ -2.664 \\ -3.226 \\ 1 \end{bmatrix},$$

$$\lambda_3 = 4.427; v_3 = \begin{bmatrix} 2.527 \\ -1.644 \\ -5.42 \\ 1 \end{bmatrix}; \ \lambda_4 = 8.66; v_4 = \begin{bmatrix} 1.493 \\ -9.108 \\ 3.644 \\ 1 \end{bmatrix}$$

Choose the eigen vectors corresponding to top two eigen values. Here λ_3 and λ_4.

Suppose we decide to reduce 4-dim data to 2 dim data:

$$A_{2\times 4} = \begin{bmatrix} 1.493 & -9.108 & 3.644 & 1 \\ 2.527 & -1.644 & -5.42 & 1 \end{bmatrix}$$

$$y_i = A(x_i - m_x). \qquad \text{Here } m_x = 4.85$$

For Arun

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1.493 & -9.108 & 3.644 & 1 \\ 2.527 & -1.644 & -5.42 & 1 \end{bmatrix} \left[\begin{bmatrix} 4 \\ 8 \\ 4 \\ 3 \end{bmatrix} - 4.85 \right]$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1.493 & -9.108 & 3.644 & 1 \\ 2.527 & -1.644 & -5.42 & 1 \end{bmatrix} \begin{bmatrix} -0.85 \\ 3.15 \\ -0.85 \\ -1.85 \end{bmatrix} = \begin{bmatrix} -34.90665 \\ -4.56955 \end{bmatrix}$$

For Rajiv

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1.493 & -9.108 & 3.644 & 1 \\ 2.527 & -1.644 & -5.42 & 1 \end{bmatrix} \left[\begin{bmatrix} 2 \\ 6 \\ 9 \\ 3 \end{bmatrix} - 4.85 \right]$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1.493 & -9.108 & 3.644 & 1 \\ 2.527 & -1.644 & -5.42 & 1 \end{bmatrix} \begin{bmatrix} -2.85 \\ 1.15 \\ 4.15 \\ -1.85 \end{bmatrix}$$

$$= \begin{bmatrix} -1.45779 \\ -33.43555 \end{bmatrix}$$

For Geeta

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1.493 & -9.108 & 3.644 & 1 \\ 2.527 & -1.644 & -5.42 & 1 \end{bmatrix} \begin{bmatrix} \begin{bmatrix} 6 \\ 1 \\ 7 \\ 4 \end{bmatrix} - 4.85 \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1.493 & -9.108 & 3.644 & 1 \\ 2.527 & -1.644 & -5.42 & 1 \end{bmatrix} \begin{bmatrix} 1.15 \\ -3.85 \\ 2.15 \\ -0.85 \end{bmatrix}$$

$$= \begin{bmatrix} 43.76735 \\ -3.26755 \end{bmatrix}$$

For Ramesh

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1.493 & -9.108 & 3.644 & 1 \\ 2.527 & -1.644 & -5.42 & 1 \end{bmatrix} \left[\begin{bmatrix} 5 \\ 9 \\ 6 \\ 3 \end{bmatrix} - 4.85 \right]$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1.493 & -9.108 & 3.644 & 1 \\ 2.527 & -1.644 & -5.42 & 1 \end{bmatrix} \begin{bmatrix} 0.15 \\ 4.15 \\ 1.15 \\ -1.85 \end{bmatrix}$$

$$= \begin{bmatrix} -35.23365 \\ -14.52655 \end{bmatrix}$$

For Sunder

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1.493 & -9.108 & 3.644 & 1 \\ 2.527 & -1.644 & -5.42 & 1 \end{bmatrix} \left[\begin{bmatrix} 4 \\ 6 \\ 3 \\ 4 \end{bmatrix} - 4.85 \right]$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1.493 & -9.108 & 3.644 & 1 \\ 2.527 & -1.644 & -5.42 & 1 \end{bmatrix} \begin{bmatrix} -0.85 \\ 1.15 \\ -1.85 \\ -0.85 \end{bmatrix}$$

$$= \begin{bmatrix} -19.33465 \\ 5.13845 \end{bmatrix}$$

The transformed vectors in 2-dim are:

Arun	−34.90665	−4.56955
Rajiv	−1.45779	−33.43555
Geeta	43.76735	−3.26755
Ramesh	−35.23365	−14.52655
Sunder	−19.33465	5.13845

The original data vectors with certain error can be recovered by:

$$\acute{x}_i = A^T y_i + m_x$$

he mean square error (mse) between the original and reconstructed vectors is the sum of the eigenvalues whose corresponding eigenvectors are not used in the transformation matrix ***A***.

$$e_{mse} = \sum_{j=k+1}^{n} \lambda_j$$

Error = sum of eigen values NOT used =

$$\lambda_1 + \lambda_2 = 0.032 + 1.1041 = 1.1361$$

Generalized PCA Steps

1. Start with N d $-$dimensional data vectors, $\boldsymbol{x}_i, i = 1, 2, \ldots N$, and find the eigenvalues and eigenvectors of the sample covariance matrix of size $d\ x\ d$ using the given data.
2. Select the top k eigenvalues, $k \leq d$, and use the corresponding eigenvectors to define the linear transformation matrix A of size $k \times d$ for transforming original features into the new space.

3. Obtain the transformed vectors, $\boldsymbol{y}_i, i = 1, 2, \ldots N$, using

$$y_i = A(x_1 - m_x)$$

The transformation involves first shifting the origin of the original feature space using the mean of the input vectors, m_x.

4. These transformed vectors are used for visualization and building predictive model.
5. The original data vectors can be recovered with certain error by:

$$\hat{x}_i = A^t y_i + m_x$$

6. The mean square error (mse) between the original and reconstructed vectors is the sum of the eigenvalues whose corresponding eigenvectors are **not** used in the transformation matrix ***A***.

$$error = \sum_{j=k+1}^{d} \lambda_j$$

7. Another way to look at the performance of PCA is by calculating the percentage variability, *P*, captured by eigenvectors corresponding to top k eigenvalues.

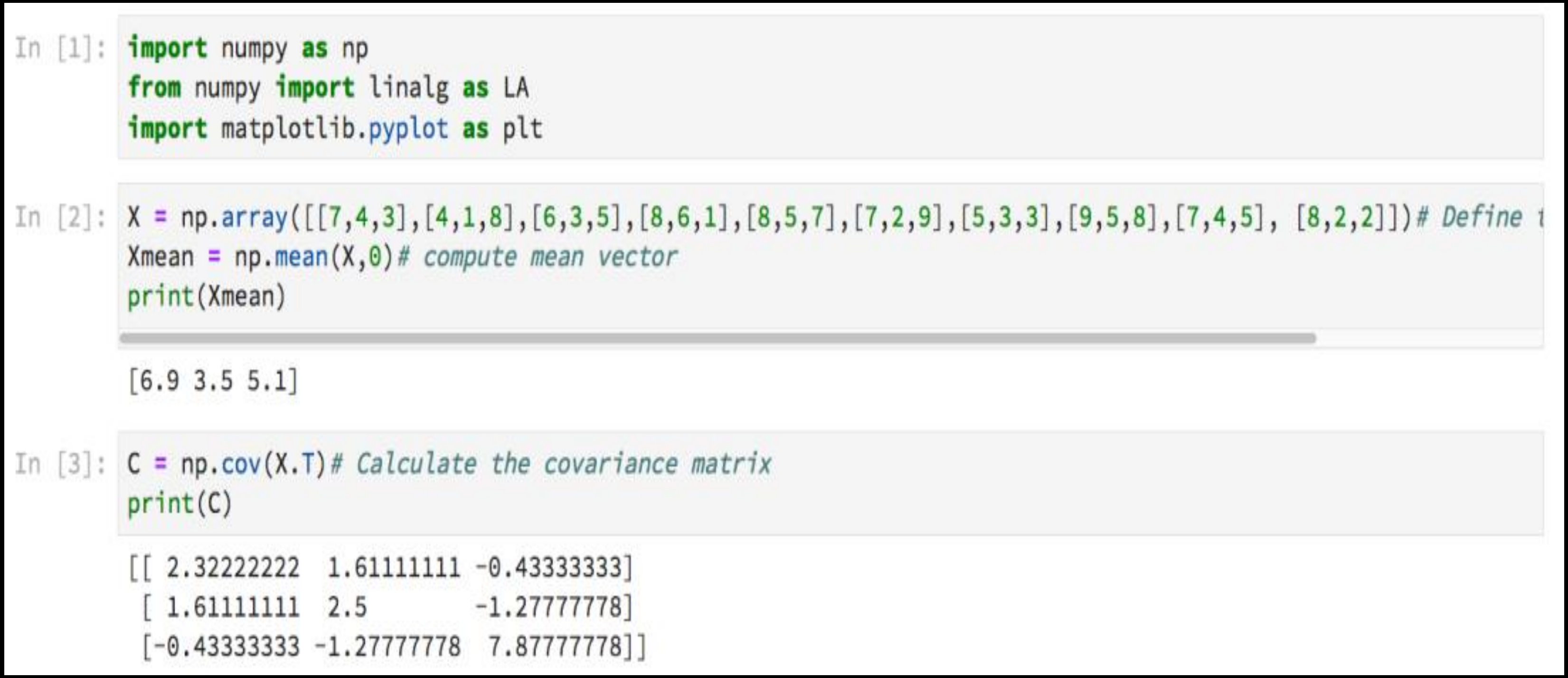
$$P = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{d} \lambda_j}$$

Ex 3 (using python):

Let the marks of 10 students be given:

Student	Maths	English	Hindi
1	7	4	3
2	4	1	8
3	6	3	5
4	8	6	1
5	8	5	7
6	7	2	9
7	5	3	3
8	9	5	8
9	7	4	5
10	8	2	2

1. Calculate mean and covariance matrix

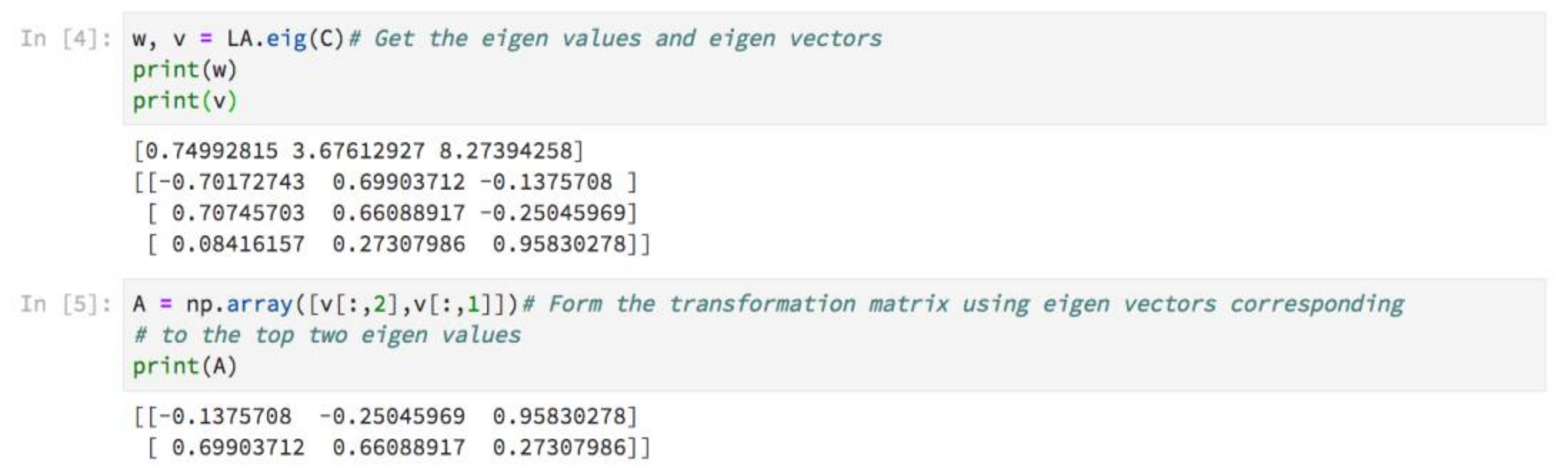


```
In [1]: import numpy as np
        from numpy import linalg as LA
        import matplotlib.pyplot as plt
```

```
In [2]: X = np.array([[7,4,3],[4,1,8],[6,3,5],[8,6,1],[8,5,7],[7,2,9],[5,3,3],[9,5,8],[7,4,5], [8,2,2]])# Define t
        Xmean = np.mean(X,0)# compute mean vector
        print(Xmean)

        [6.9 3.5 5.1]
```

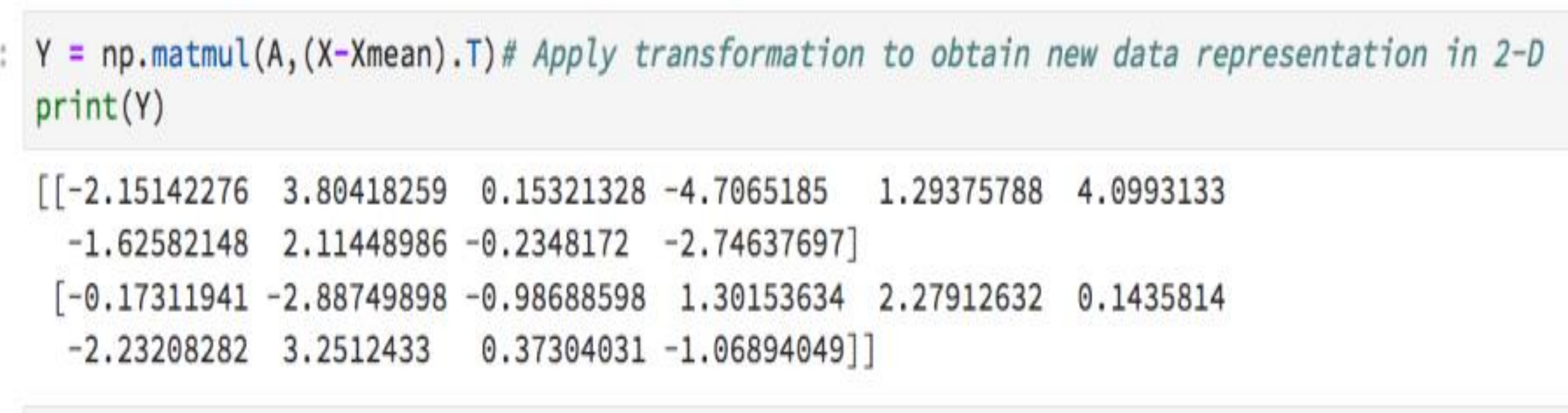
```
In [3]: C = np.cov(X.T)# Calculate the covariance matrix
        print(C)

        [[ 2.32222222  1.61111111 -0.43333333]
         [ 1.61111111  2.5        -1.27777778]
         [-0.43333333 -1.27777778  7.87777778]]
```

2. Compute eigenvalues and eigenvectors. To reduce the data to two dimensions, form the transformation matrix ***A*** using the eigenvectors of top two eigenvalues.



```
In [4]: w, v = LA.eig(C)# Get the eigen values and eigen vectors
        print(w)
        print(v)

        [0.74992815 3.67612927 8.27394258]
        [[-0.70172743  0.69903712 -0.1375708 ]
         [ 0.70745703  0.66088917 -0.25045969]
         [ 0.08416157  0.27307986  0.95830278]]

In [5]: A = np.array([v[:,2],v[:,1]])# Form the transformation matrix using eigen vectors corresponding
        # to the top two eigen values
        print(A)

        [[-0.1375708  -0.25045969  0.95830278]
         [ 0.69903712  0.66088917  0.27307986]]
```

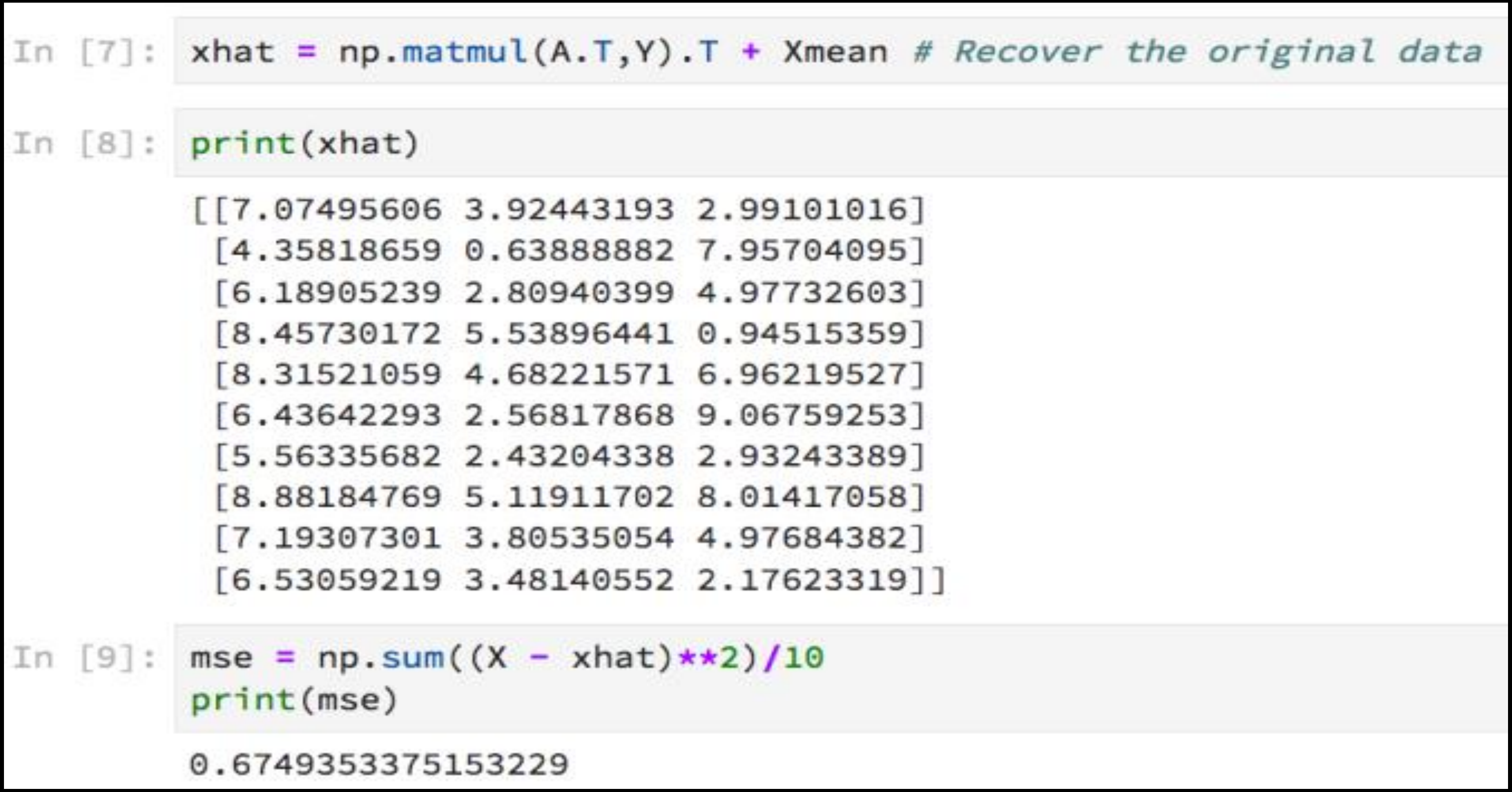
With the calculated $\boldsymbol{A}$ matrix, transform the input vectors to obtain vectors in two dimensions.



```
Y = np.matmul(A,(X-Xmean).T)# Apply transformation to obtain new data representation in 2-D
print(Y)
```

```
[[-2.15142276  3.80418259  0.15321328 -4.7065185   1.29375788  4.0993133
  -1.62582148  2.11448986 -0.2348172  -2.74637697]
 [-0.17311941 -2.88749898 -0.98688598  1.30153634  2.27912632  0.1435814
  -2.23208282  3.2512433   0.37304031 -1.06894049]]
```

Reconstruct the original vectors and compute the mean square error between the original 3-dim vectors and the reconstructed 3-dim vectors.



```
In [7]: xhat = np.matmul(A.T,Y).T + Xmean # Recover the original data
```

```
In [8]: print(xhat)
```

```
[[7.07495606 3.92443193 2.99101016]
 [4.35818659 0.63888882 7.95704095]
 [6.18905239 2.80940399 4.97732603]
 [8.45730172 5.53896441 0.94515359]
 [8.31521059 4.68221571 6.96219527]
 [6.43642293 2.56817868 9.06759253]
 [5.56335682 2.43204338 2.93243389]
 [8.88184769 5.11911702 8.01417058]
 [7.19307301 3.80535054 4.97684382]
 [6.53059219 3.48140552 2.17623319]]
```

```
In [9]: mse = np.sum((X - xhat)**2)/10
        print(mse)
```

```
0.6749353375153229
```

The calculated mean square error is not equal to the smallest eigenvalue (0.74992815) as expected. Why?

The formula used in calculating the covariance matrix assumes the number of examples, N, to be large.

But here, the number of examples is only 10.

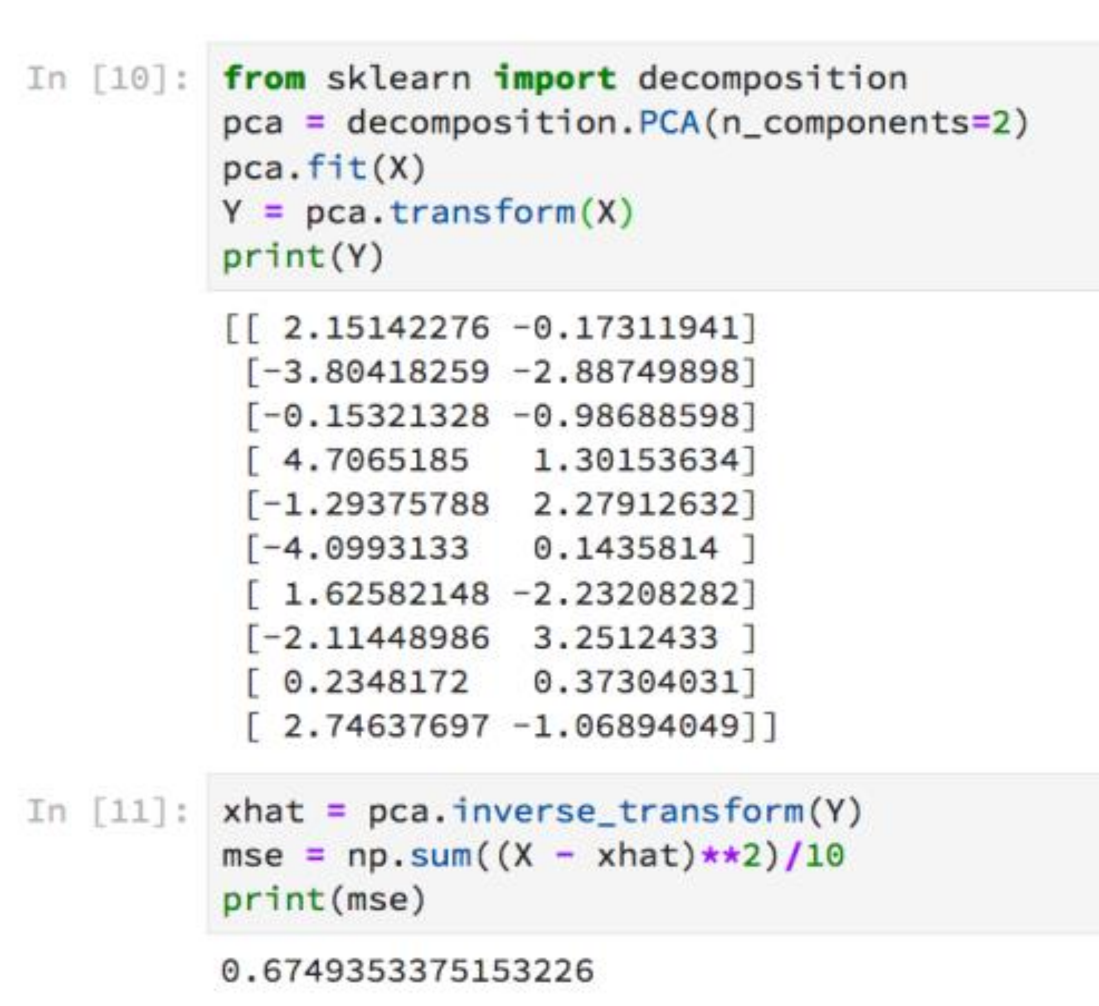
Thus multiply the mse value by N/(N-1), known as the small sample correction.

Then, the result identical to the smallest eigenvalue.

Are obtained.

As N becomes large, the ratio N/(N-1) approaches unity and no such correction is required.

In practice, the PCA can be easily done using the scikit-learn implementation as shown below.



```
In [10]: from sklearn import decomposition
         pca = decomposition.PCA(n_components=2)
         pca.fit(X)
         Y = pca.transform(X)
         print(Y)

         [[ 2.15142276 -0.17311941]
          [-3.80418259 -2.88749898]
          [-0.15321328 -0.98688598]
          [ 4.7065185   1.30153634]
          [-1.29375788  2.27912632]
          [-4.0993133   0.1435814 ]
          [ 1.62582148 -2.23208282]
          [-2.11448986  3.2512433 ]
          [ 0.2348172   0.37304031]
          [ 2.74637697 -1.06894049]]

In [11]: xhat = pca.inverse_transform(Y)
         mse = np.sum((X - xhat)**2)/10
         print(mse)

         0.6749353375153226
```

Some tips

- Instead of performing PCA using the covariance matrix, we can also use the **correlation matrix**, which has a built-in normalization of features and thus the data normalization is not needed.
- Eigenvalues and eigenvectors are typically calculated by the singular value decomposion (SVD) method of matrix factorization. Thus, PCA and SVD are often viewed the same. But you should remember that the starting point for PCA is a collection of data vectors that are needed to compute sample covariance/correlation matrices to perform eigenvector decomposition which is often done by SVD.

Differentiation of Univariate Functions

Def: The **difference quotient** computes the slope of the secant line through two points on the graph of f

$$\frac{\delta y}{\delta x} = \frac{f(x + \delta x) - f(x)}{\delta x}$$

Def: For $h > 0$ the **derivative** of f at x is defined as the limit

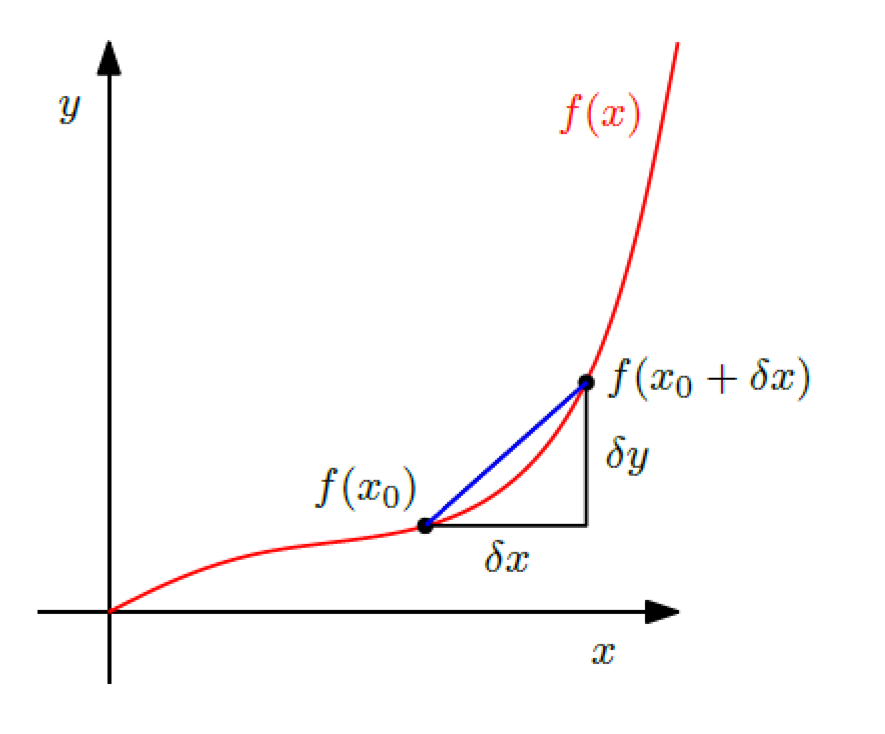
$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

and the secant becomes a tangent.

Note: The derivative of f points in the direction of steepest ascent of f.

Geometric representation

The average incline of a function f between x_0 and $x_0 + \delta x$ is the incline of the secant (blue) through $f(x_0)$ and $f(x_0 + \delta x)$ and is given by ${}^{\delta y}/_{\delta x}$



Ex: Derivative of a Polynomial

Derivative of polynomial $f(X) = x^n, n \in \mathbb{N}$ is nx^{n-1}.

$$\frac{df}{dx} = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = \lim_{h\to 0}\frac{(x+h)^n - x^n}{h}$$

$$= \lim_{h\to 0}\frac{\sum_{i=0}^{n}\binom{n}{i}x^{n-i}h^i - x^n}{h}$$

$$= \lim_{h\to 0}\frac{\sum_{i=1}^{n}\binom{n}{i}x^{n-i}h^i}{h}$$

$$= \lim_{h\to 0}\sum_{i=1}^{n}\binom{n}{i}x^{n-i}h^{i-1}$$

$$= \lim_{h\to 0}\binom{n}{1}x^{n-1} + \sum_{i=2}^{n}\binom{n}{i}\,x^{n-i}h^{i-1}$$

$$= \frac{n!}{1!(n-1)!}x^{n-1} = nx^{n-1}$$

Def: The **Taylor polynomial** of degree n of $f: \mathbb{R} \rightarrow \mathbb{R}$, at x_0. is defined as:

$$T_n(x) = \sum_{k=1}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

Where $f^{(k)}(x_0)$ is the k^{th} derivative of f at x_0 (assuming it exists) and $\frac{f^{(k)}(x_0)}{k!}$ are the coefficients of the polynomial.

Def: For a smooth function $f \in \mathbb{C}^\infty, f: \mathbb{R} \rightarrow \mathbb{R}$, the **Taylor series** of f at x_0 is defined as:

$$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

For $x = x_0$, we get the **Maclaurin series** as a special instance of the Taylor series. If $f(x) = T_\infty(x)$, then f is called **analytic**, where $f \in \mathbb{C}^\infty$ means f is continuously differentiable infinitely many times.

Ex: Consider a polynomial $f(x) = x^4$, find T_6 evaluated at $x_0 = 1$.

We compute $f^{(k)}(1)$ for $k = 1, 2, \dots 6$

$$f(1) = 1;\ f'(1) = 4;\ f^2(1) = 12;$$
$$f^3(1) = 24;\ f^4(1) = 24;\ f^5(1) = 0;\ f^6(1) = 0.$$

The desired Taylor polynomial is:

$$T_6(x) = \sum_{k=1}^{6} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$
$$= 1 + 4(x-1) + 6(x-1)^2 + 4(x-1)^3 + (x-1)^4 + 0$$
$$= x^4 = f(x)$$

Ex: Consider $f(x) = \sin x + \cos x \in \mathbb{C}^\infty$

Find the Taylor series expansion of f at $x_0 = 0$. which is the Maclaurin series expansion of f.

$$f(0) = \sin 0 + \cos 0 = 1$$

$$f'(0) = \cos 0 - \sin 0 = 1$$

$$f''(0) = -\sin 0 - \cos 0 = -1$$

$$f^{(3)}(0) = -\cos 0 + \sin 0 = -1$$

$$f^{(4)}(0) = \sin 0 + \cos 0 = 1 = f(0)$$

….

$$f^{(k+4)}(0) = f^{(k)}(0)$$

So, $T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$

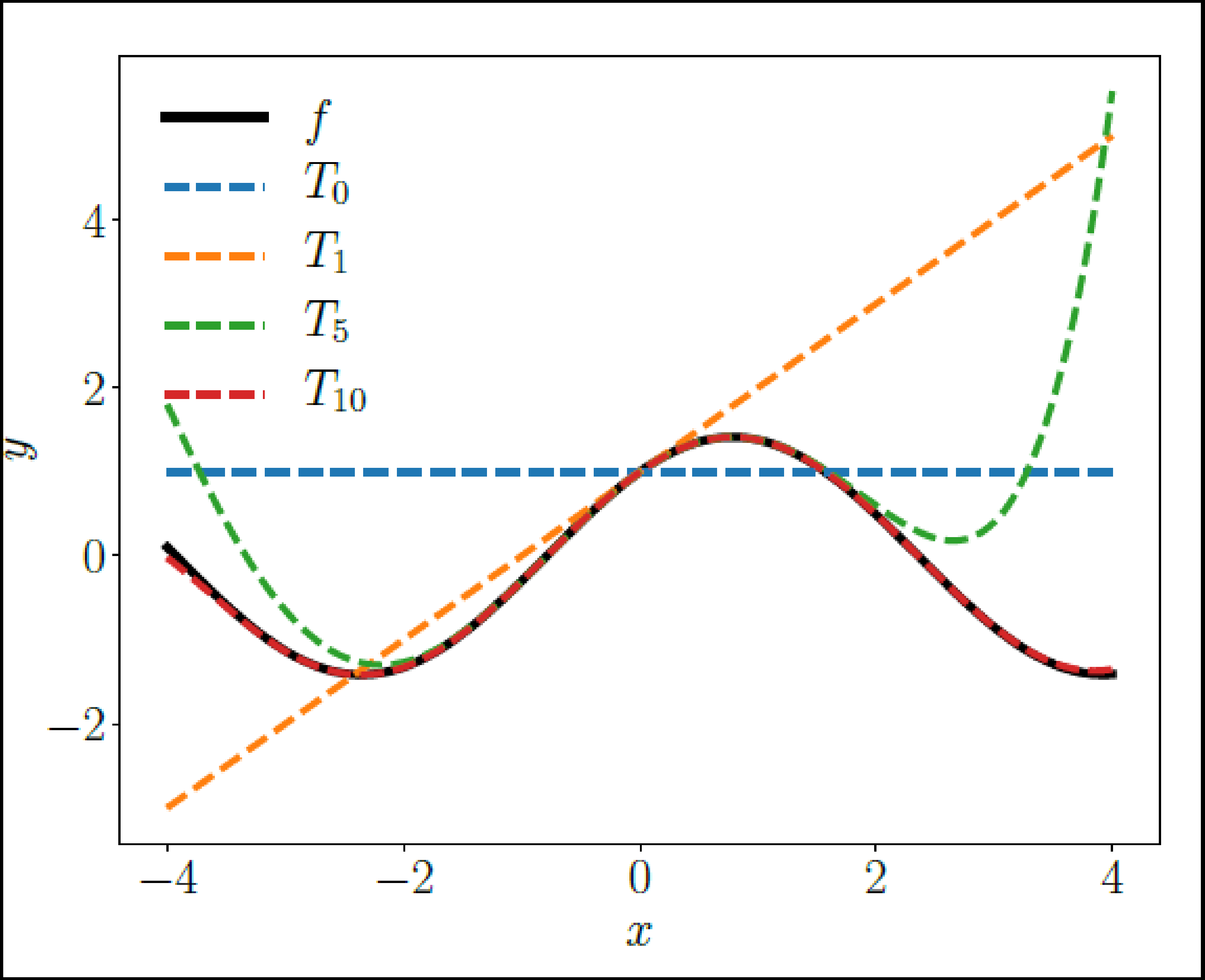
$$= 1 + x - \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 + \ \dots$$

$$= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 \mp \cdots \ x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 \mp \cdots$$

$$= \sum_{k=0}^{\infty}(-1)^k \frac{1}{(2k)!}x^{2k} + \sum_{k=0}^{\infty}(-1)^k \frac{1}{(2k+1)!}x^{2k+1}$$

$$= \cos(x) + \sin(x)$$

first Taylor polynomials T_n for $n = 0,1,5,10.$



Note: A Taylor series is a special case of a power series

$$f(x) = \sum_{k=0}^{\infty} a_k (x - c_k)^k$$

Where a_k are coefficients and c is a constant

Differentiation Rules

Product rule: $(f(x)g(x))' = f'(x)g(x) + g'(x)f(x)$

Quotient rule: $\left(\frac{f(X)}{g(x)}\right)' = \frac{f'(x)g(x) - g'(x)f(x)}{(g(x))^2}$

Sum rule: $(f(x) + g(x))' = f'(x) + g'(x)$

Chain rule: $(g(f(x)))' = (g \circ f)'(x) = g'(f(x))f'(x)$

Partial Differentiation and Gradients

Let f be a function of many variables, i.e. $\boldsymbol{x} \in \mathbb{R}^n$.

The gradient of the function f with respect to $\boldsymbol{x}$ is obtained by varying one variable at a time and keeping the others constant. The gradient is the collection of these partial derivatives.

Def: For a function $f: \mathbb{R}^n \to \mathbb{R}, x \longmapsto f(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n$ of n variables $x_1, x_2, \ldots\ x_n$ the partial derivatives are defined as:

$$\frac{\partial f}{\partial x_1} = \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots\ x_n) - f(\boldsymbol{x})}{h}$$

$$\vdots$$

$$\frac{\partial f}{\partial x_n} = \lim_{h \to 0} \frac{f(x_1, x_2, \ldots\ x_n + h) - f(\boldsymbol{x})}{h}$$

And collect them in a row vector:

$$\nabla_{\boldsymbol{x}} f = grad\ f = \frac{df}{dx} = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1}\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \cdots \frac{\partial f(\boldsymbol{x})}{\partial x_n}\right] \in \mathbb{R}^{1 \times n}$$

Where n is the number of variables and $\boldsymbol{x} = [x_1, x_2, \ldots\ x_n]^T \in \mathbb{R}^n$. The row vector is called the **gradient or jacobian** of f.

Ex: Find Partial Derivative using chain rule of

$$f(x, y) = (x + 2y^3)^2$$

$$\frac{\partial f}{\partial x} = 2(x + 2y^3)\frac{\partial}{\partial x}(x + 2y^3) = 2(x + 2y^3)$$

$$\frac{\partial f}{\partial y} = 2(x + 2y^3)\frac{\partial}{\partial y}(x + 2y^3) = 12(x + 2y^3)y^2$$

Gradient vector is: $[2(x + 2y^3) \quad 12(x + 2y^3)y^2]$

Basic Rules of Partial Differentiation

Product rule: $\frac{\partial}{\partial x}\left(f(x)g(x)\right) = \frac{\partial}{\partial x}g(x) + f(x)\frac{\partial g}{\partial x}$

Sum rule: $\frac{\partial}{\partial x}\left(f(x) + g(x)\right) = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial x}$

Chain rule: $\frac{\partial}{\partial x}\left((g \circ f)(x)\right) = \frac{\partial}{\partial x}g\left(f(x)\right) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x}$

Consider a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ of two variables x_1, x_2. Also, x_1 and x_2 are themselves functions of t. To compute the gradient of f with respect to t, we need to apply the chain rule for multivariate functions as:

$$\frac{df}{dt} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}$$

Where d is derivate and ∂ are partial derivatives.

Ex: If $f(x_1, x_2) = x_1^2 + 2x_2$, such that $x_1 = \sin t$ and $x_2 = \cos t$, then the partial derivatives are:

$$\frac{df}{dt} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}$$

$$= 2\sin t\frac{\partial \sin t}{\partial t} + 2\frac{\partial\, cos\, t}{\partial t}$$

$$= 2\sin t\cos t - 2\sin t$$

Ex: If $f(x_1, x_2)$ is a function of x_1 and x_2 , where $x_1(s,t)$ and $x_2(s,t)$ are themselves function of two variables s and t, then the partial derivatives are:

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s}$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}$$

The gradient is obtained by matrix multiplication:

$$\frac{df}{d(s,t)} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial(s,t)} = \begin{bmatrix}\frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2}\end{bmatrix}\begin{bmatrix}\frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t}\end{bmatrix}$$

The only nonzero third-order partial derivative is:

$$\frac{\partial^3 f}{\partial y^3} = 6 \Longrightarrow \frac{\partial^3 f}{\partial y^3}(1,2) = 6$$

Higher-order derivatives and the mixed derivatives of degree 3 vanish, such that:

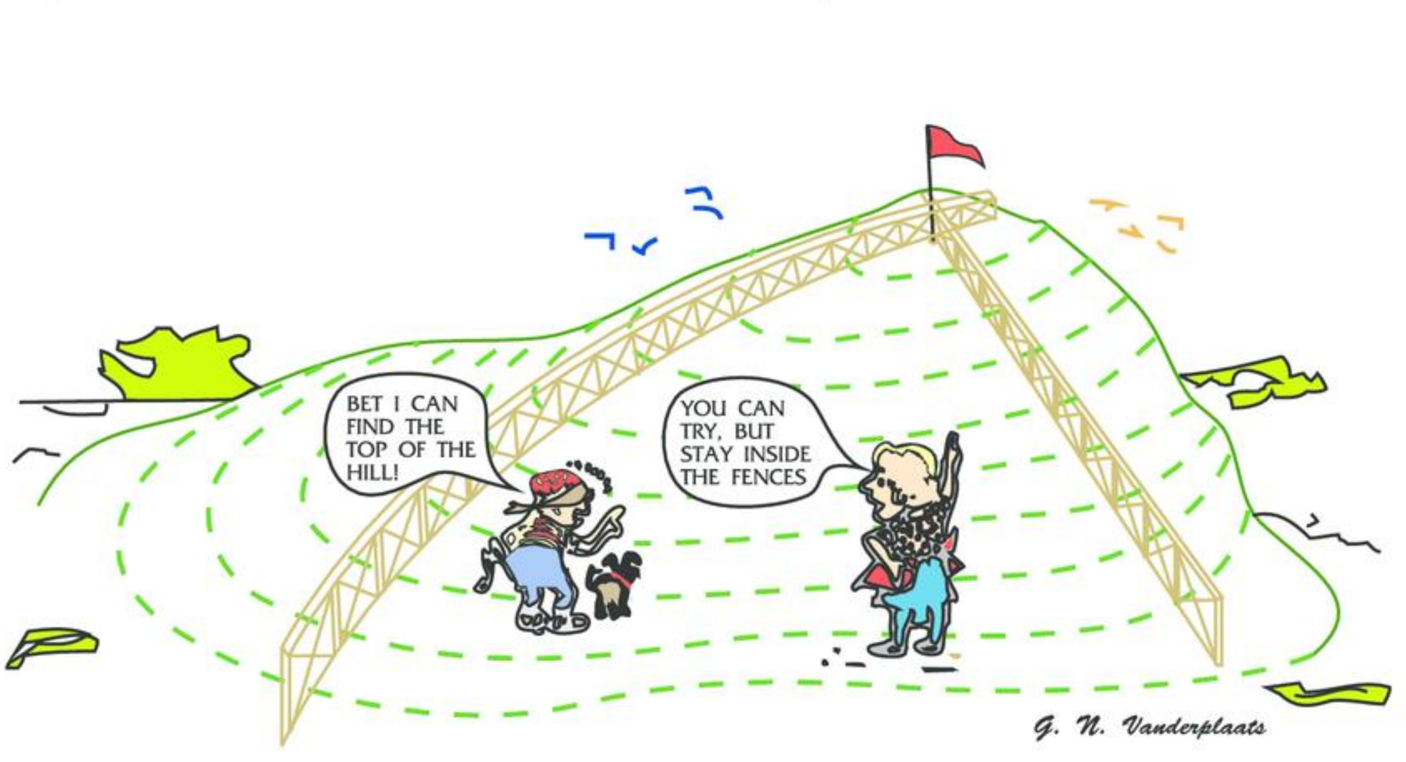
$D^3_{x,y} f[:,:,1] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ and $D^3_{x,y} f[:,:,2] = \begin{bmatrix} 0 & 0 \\ 0 & 6 \end{bmatrix}$

And $\frac{D^3_{x,y} f(1,2)}{3!} \delta^3 = (y-2)^3$

What is optimization ?

Choosing “the best” amongst multiple options.

tion, and the constraints are the fences which he must stay inside of.

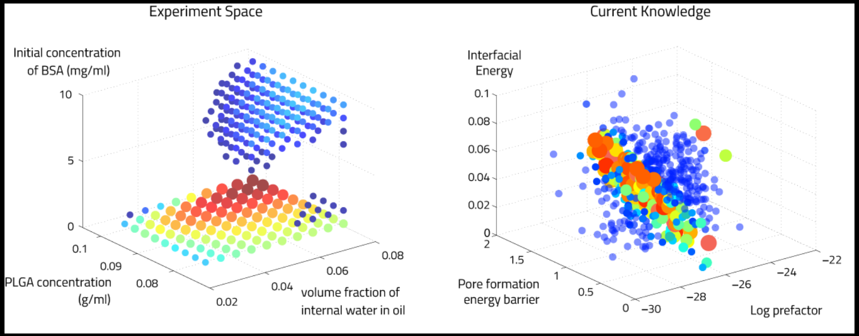

BET I CAN FIND THE TOP OF THE HILL!
YOU CAN TRY, BUT STAY INSIDE THE FENCES
G. N. Vanderplaats

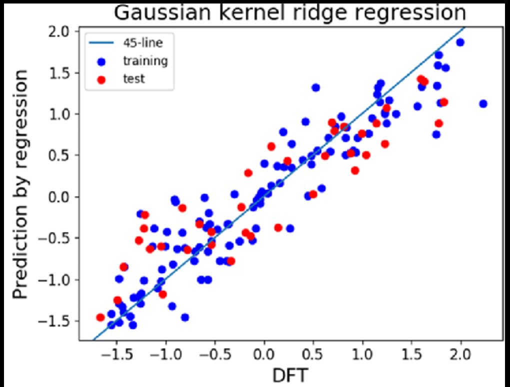
Choosing “the best” amongst many alternatives.

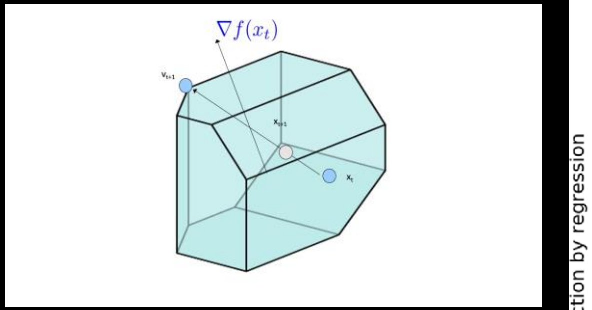
Examples:

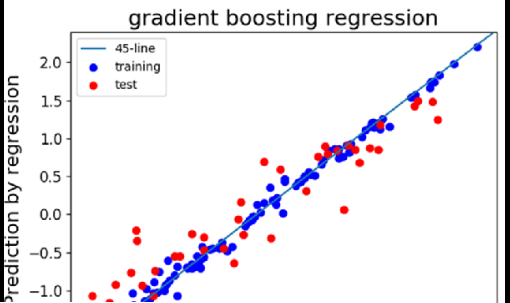
- Optimal (efficient) design of machines
- Minimize cost
- Maximize reliability
- Least square estimators
- Optimization weights in NN

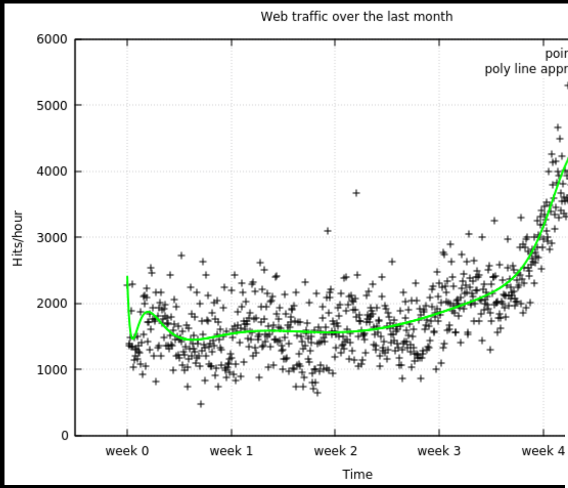
Optimization in Machine Learning

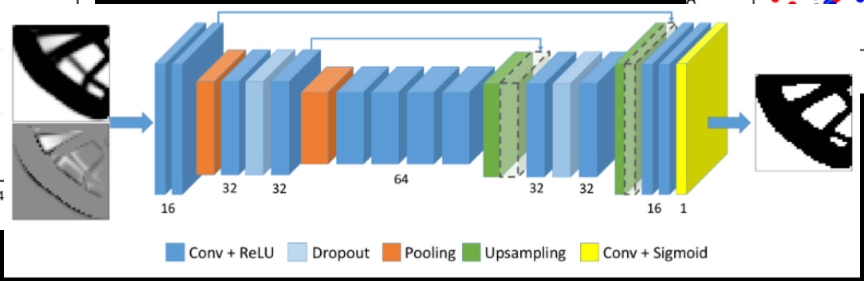










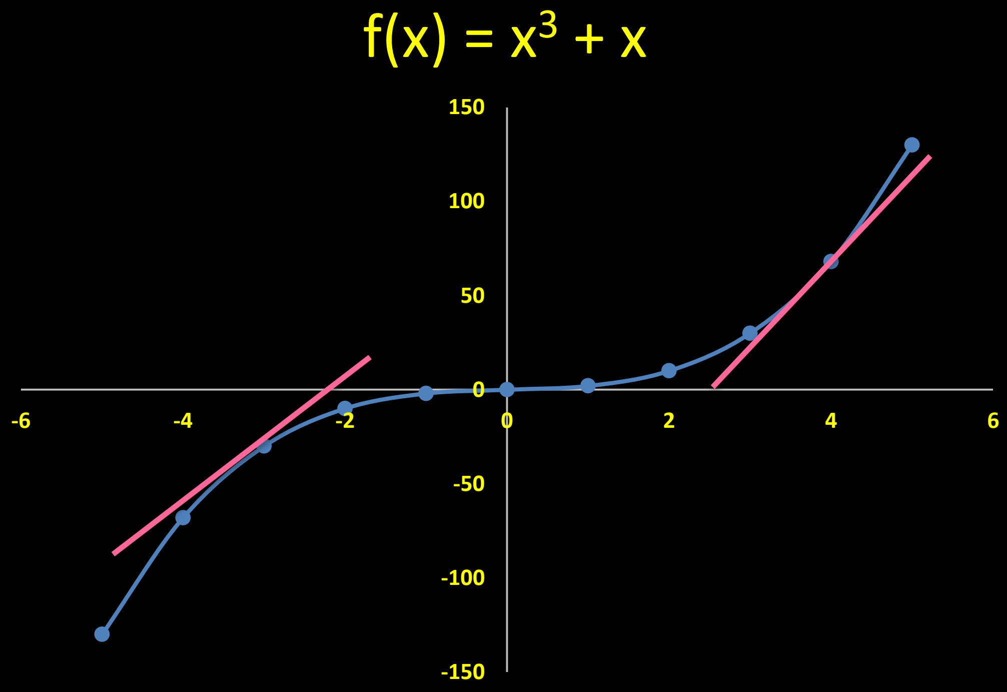


- Classification
- Feature selection
- Regression Analysis
- Topology optimization
- Training Neural Nets by optimizing weights of Neural Network

Ex 1: $f(x) = x^3 + x$ over any interval [a, b]

Slope of tangent = $f'(x) = 3x^2 + 1$ is always > 0.

So, f(x) is an increasing function for all x.



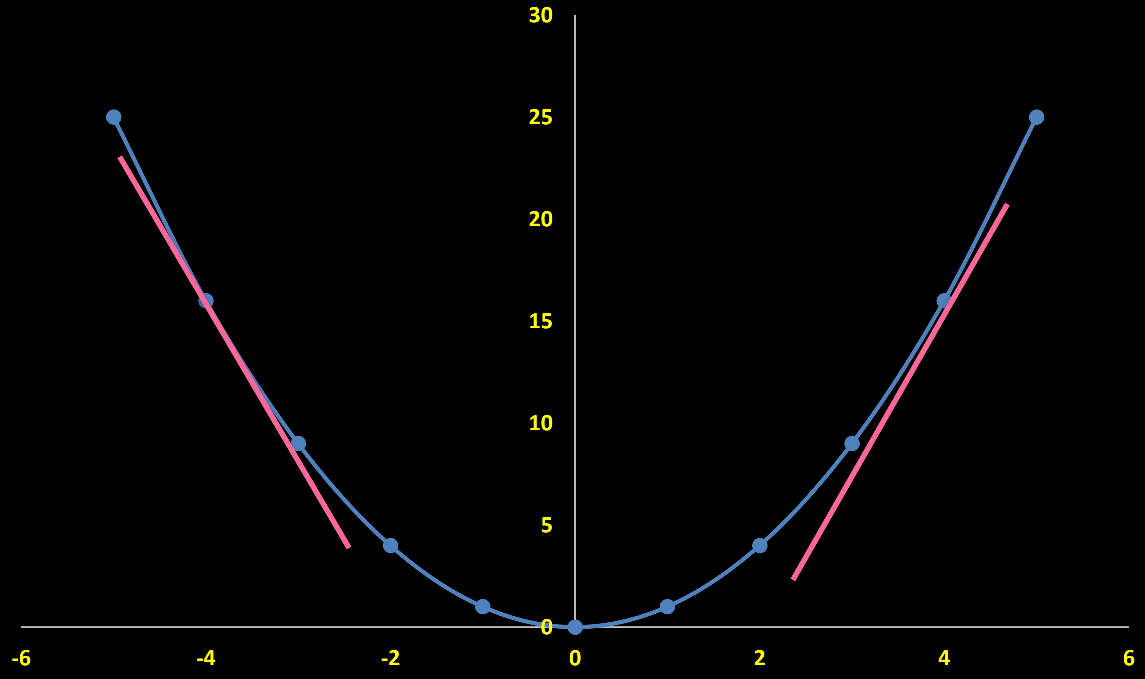
Ex 2: $f(x) = x^2$

When $x > 0$, Slope of tangent = $f'(x) = 2x > 0$.

So, $f(x)$ is an <u>increasing</u> function for all $x > 0$.

When $x < 0$, Slope of tangent = $f'(x) = -2x < 0$.

So, $f(x)$ is an <u>decreasing</u> function for all $x < 0$.



Def: Stationary point is a point where derivative is zero.

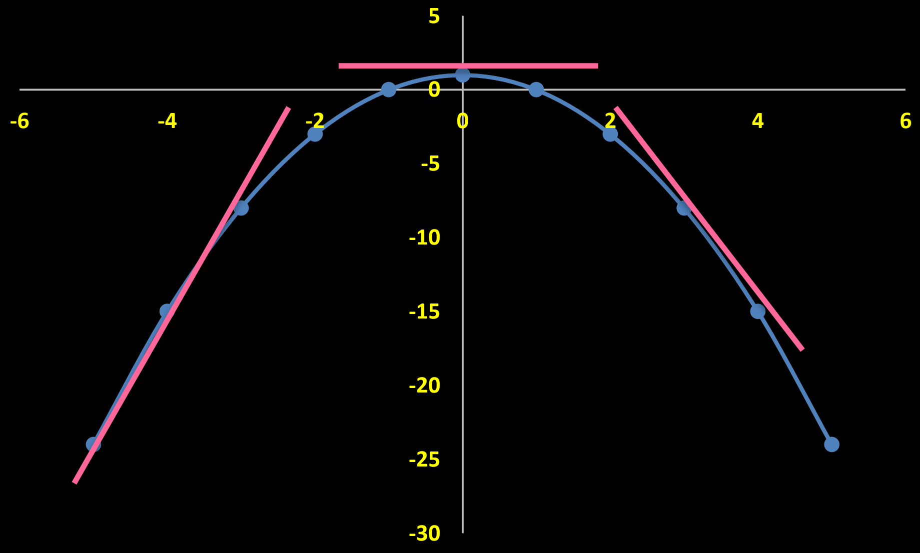
Stationary point can be either of the following:

- Local minima or relative minima
- Local maxima or relative maxima
- Point of inflection

Ex 3: f(x) = 1 – x^2

f’(x) = 0 gives –2x = 0. x = 0 is a maxima

$$f'(x) = \begin{cases} > 0 & for \quad x < 0 \\ = 0 & for \quad x = 0 \\ < 0 & for \quad x > 0 \end{cases}$$



Ex 4: $f(x) = 2x^3 – 6x^2 -48x + 24$

Putting f’(x) = 0 gives stationary points:

$6x^2 – 12x – 48 = 0$ or $x^2 – 2x – 8 = 0$

Gives x = 4 and x = – 2 as two stationary points.

$f''(x) = 2x – 2$

Putting f’’(4) = 6 > 0.

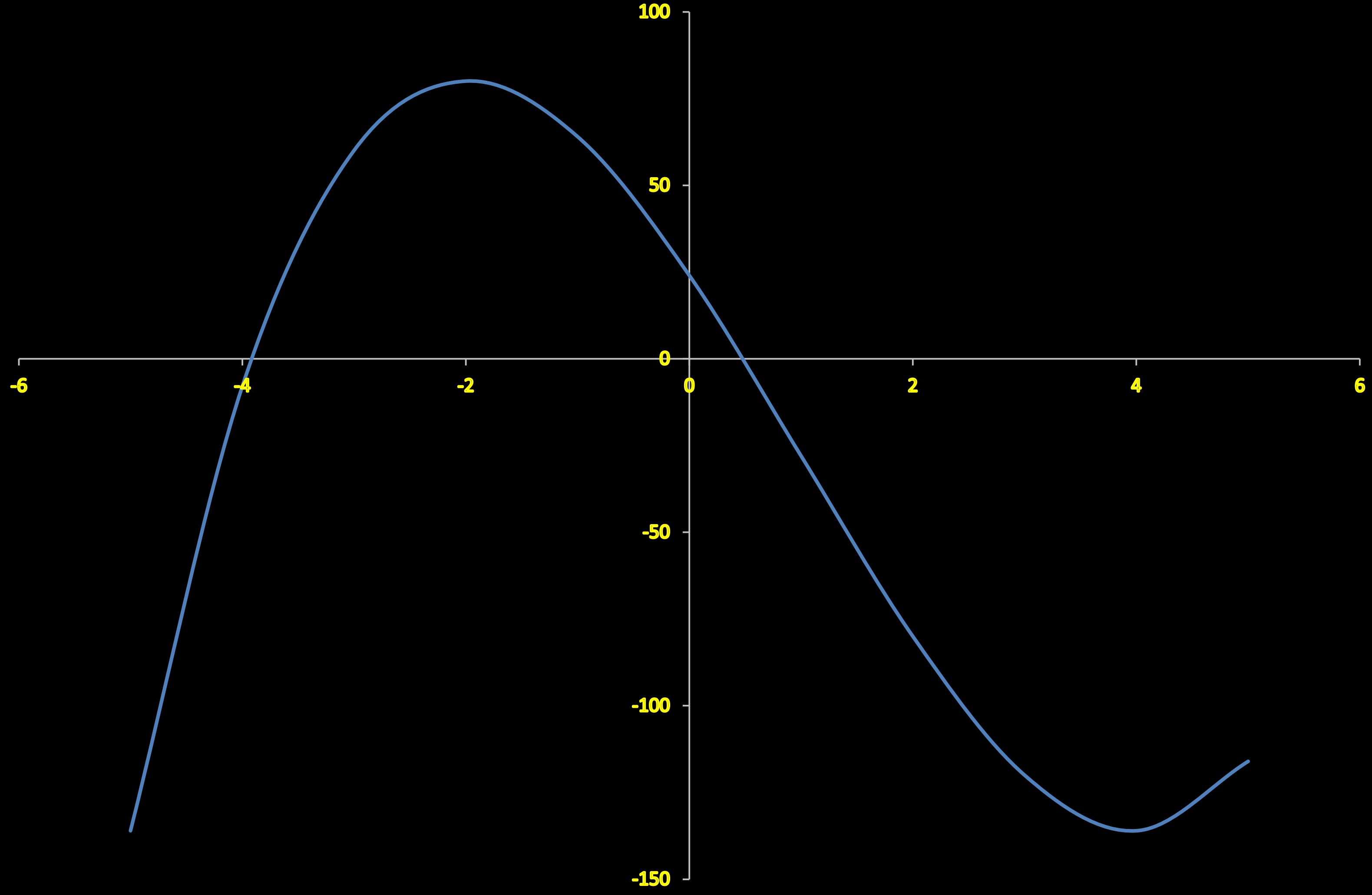
Minima at 4 with f(4) = –136

Putting f’’(–2) = –6 < 0. Maxima at –2.

Checking f(3) = –30 and f(5) = 42

And f(–1) = – 30 and f(– 3) = 42

$f(x) = 2x^3 - 6x^2 - 48x + 24$



Ex 5: $f(x) = x^4 + 7x^3 + 5x^2 – 17x + 3$

$f'(x) = 0$ gives $4x^3 + 21x^2 + 10x – 17x = 0$

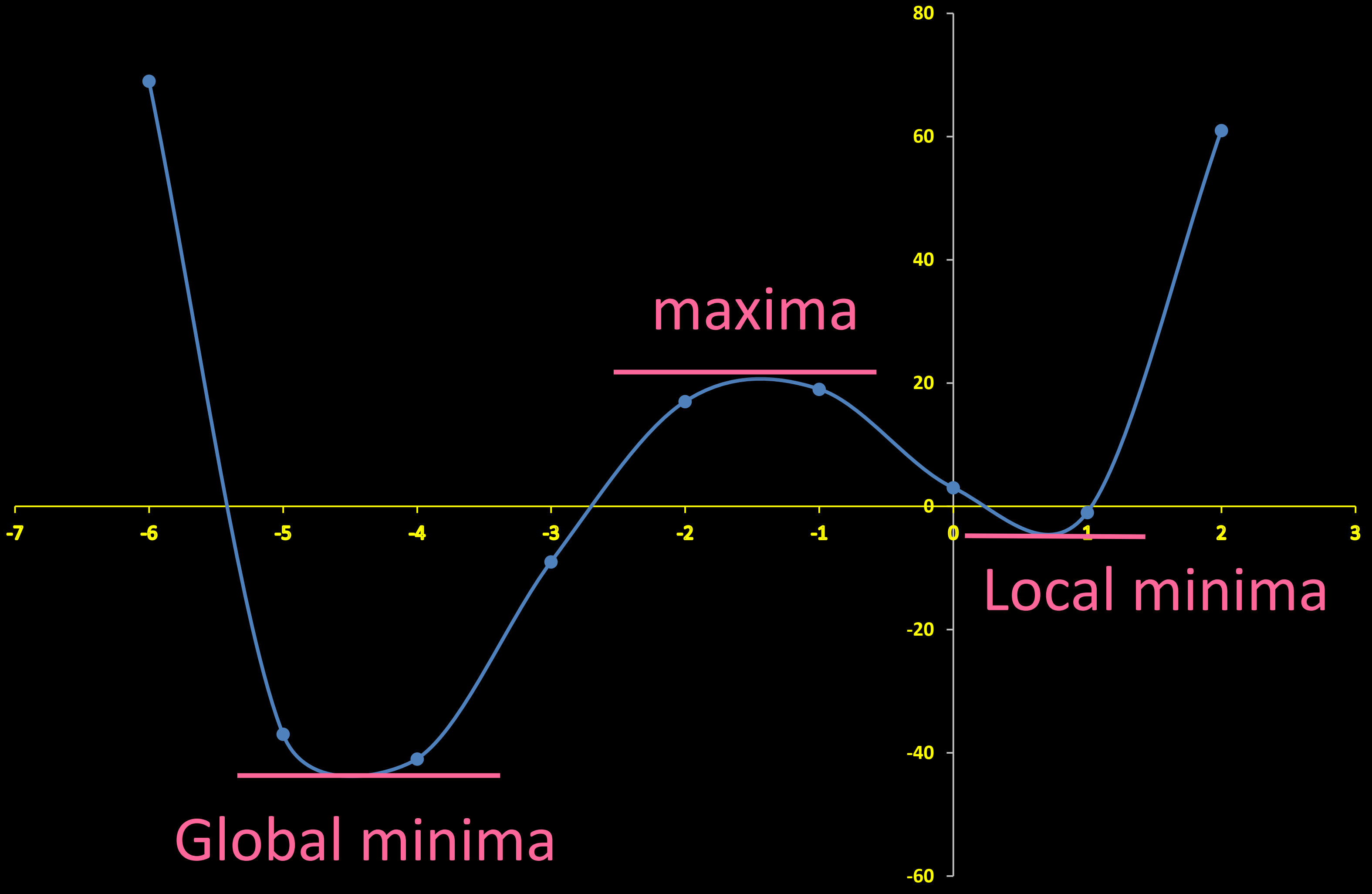
This has three roots:

$x = – 4.5$

$x = – 1.4$

$x = 0.7$

$f(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$

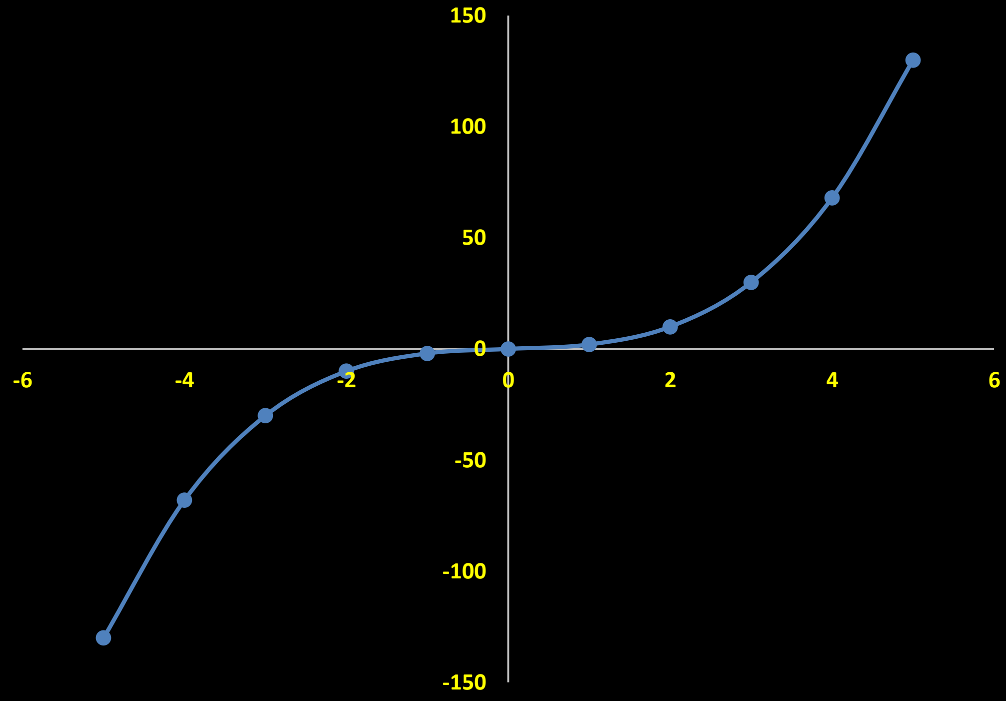


Ex 1 (revisited): $f(x) = x^3 + x$

$f'(x) = 0$ gives $3x^2 + 1 = 0$, x is not real.

Also $f'(0) = 0$.

x = 0 is a point of inflection



NON LINEAR OPTIMIZATION PROBLEM

Minimize (Maximize) f (X)

where $f: R^n \rightarrow R$, $X =(x_1, x_2, x_3 \ldots x_n)$

s.t. $X \in S \subseteq R^n$, where S is defined by

$g_k(X) \geq 0, k = 0, 1, 2, \ldots m$

$h_j (X) = 0, j = 0, 1, 2, \ldots l$

$lower_i \leq x_i \leq =upper_i$ where $i=1,2, \ldots n$

LOCAL OPTIMAL SOLUTIONS

Let D be the set of feasible values of X satisfying all the constraints.

If $\bar{X} \in D$ and $\exists$ an $N_\varepsilon(\bar{X})$ around $\bar{X}$ s.t $f(X) \geq f(\bar{X})$ for each $X \in D \cap N_\varepsilon(\bar{X})$

Then $\bar{X}$ is called a local minima.

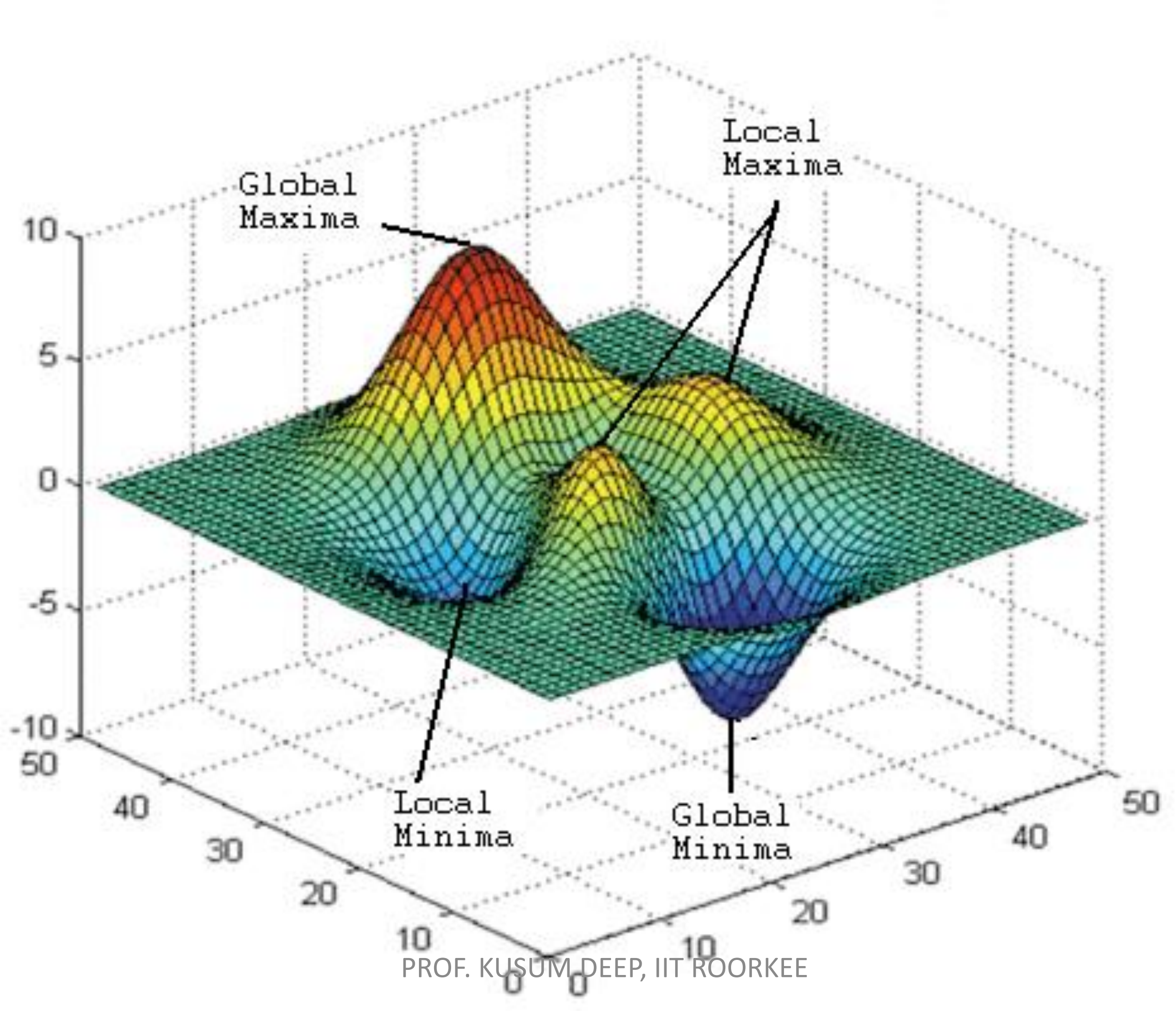
GLOBAL OPTIMAL SOLUTION

Let D be the set of feasible values of X satisfying all the constraints.

If $\bar{X} \in D$ and $f(X) \geq f(\bar{X})$ for all $X \in D$

Then $\bar{X}$ is called a Global minima.

Local and Global Optimal Solution



Nonlinear Constrained optimization

Consider Maximize $Z = f(\boldsymbol{X})$

Subject to $\boldsymbol{g}(\boldsymbol{X}) \leq \mathbf{0}$

Kuhn-Tucker necessary conditions to determine $\boldsymbol{X}$ and $\boldsymbol{\lambda}$ **to** be a stationary point.

$$\boldsymbol{\lambda} \geq \mathbf{0}$$

$$\boldsymbol{\nabla} f(\boldsymbol{X}) - \boldsymbol{\lambda}\boldsymbol{\nabla}\boldsymbol{g}(\boldsymbol{X}) = \mathbf{0}$$

$$\lambda_i g_i(\boldsymbol{X}) = 0, \quad i = 1,2, \ldots m$$

$$\boldsymbol{g}(\boldsymbol{X}) \leq \mathbf{0}$$

Sufficient Kuhn-Tucker conditions

	Required conditions	
Sense of optimization	Objective function	Solution space
Maximization	Concave	Convex set
Minimization	Convex	Convex set

Kuhn-Tucker necessary conditions

Consider generalized nonlinear optimization problem is:

Maximize (Minimize) $z = f(\boldsymbol{X})$

Subject to $\quad g_i(\boldsymbol{X}) \leq 0$ for $i = 1,2, \dots r$

$g_i(\boldsymbol{X}) = 0$ for $i = r + 1, \dots p$

$g_i(\boldsymbol{X}) \geq 0$ for $i = p + 1, \dots$m

Lagrange is:

$$L(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\lambda}) = f(\boldsymbol{X}) - \sum_{i=1}^{r} \lambda_i [g_i(\boldsymbol{X}) + S_i^2] - \sum_{i=r+1}^{p} \lambda_i [g_i(\boldsymbol{X}) - S_i^2] - \sum_{i=p+1}^{m} \lambda_i g_i(\boldsymbol{X})$$

Where λ_i are lagrange multiplier multipliers of $g_i(\boldsymbol{X})$

Kuhn-Tucker sufficient conditions

		Conditions required		
Sense of optimization	$f(\boldsymbol{X})$	$g_i(\boldsymbol{X})$	λ_i	
		Convex	≥ 0	$1 \leq i \leq r$
Maximization	concave	Concave	≤ 0	$r + 1 \leq i \leq p$
		linear	unrestricted	$p + 1 \leq i \leq m$
		Convex	≤ 0	$1 \leq i \leq r$
Minimization	convex	Concave	≥ 0	$r + 1 \leq i \leq p$
		Linear	unrestricted	$p + 1 \leq i \leq m$

In case of maximization, the Lagrangian is concave

In case of minimization, the Lagrangian is convex

Ex: Minimize $f(X) = x_1^2 + x_2^2 + x_3^2$

Subject to: $g_1(\boldsymbol{X}) = 2x_1 + x_2 - 5 \leq 0$

$$g_2(\boldsymbol{X}) = x_1 + x_3 - 2 \leq 0$$

$$g_3(\boldsymbol{X}) = 1 - x_1 \leq 0$$

$$g_4(\boldsymbol{X}) = 2 - x_2 \leq 0$$

$$g_5(\boldsymbol{X}) = -x_3 \leq 0$$

This is a minimization problem , so $\boldsymbol{\lambda} \leq \mathbf{0}$.

Kuhn-Tucker conditions become:

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) \leq 0$

$$[2x_1 \quad 2x_2 \quad 2x_3] - [\lambda_1 \quad \lambda_2 \quad \lambda_3 \quad \lambda_4 \quad \lambda_5]\begin{bmatrix} 2 & 1 & 0 \\ 1 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} = 0$$

$\lambda_1 g_1 = \lambda_2 g_2 = \lambda_3 g_3 = \lambda_4 g_4 = \lambda_5 g_5 = 0$

$\boldsymbol{g}(\boldsymbol{X}) \leq \mathbf{0}$

That is: $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) \leq 0$

$$2x_1 - 2\lambda_1 - \lambda_2 + \lambda_3 = 0$$

$$2x_2 - \lambda_1 + \lambda_4 = 0$$

$$2x_3 - \lambda_2 + \lambda_5 = 0$$

$$\lambda_1(2x_1 + x_2 - 5) = 0$$

$$\lambda_2(x_1 + x_3 - 2) = 0$$

$$\lambda_3(1 - x_1) = 0$$

$$\lambda_4(2 - x_2) = 0$$

$$\lambda_5(x_3) = 0$$

$$2x_1 + x_2 \leq 5$$

$$x_1 + x_3 \leq 2$$

$$1 - x_1 \leq 0$$

$$2 - x_2 \leq 0$$

$$-x_3 \leq 0$$

Solving these equations gives:

$$x_1 = 1, x_2 = 2, x_3 = 0,$$

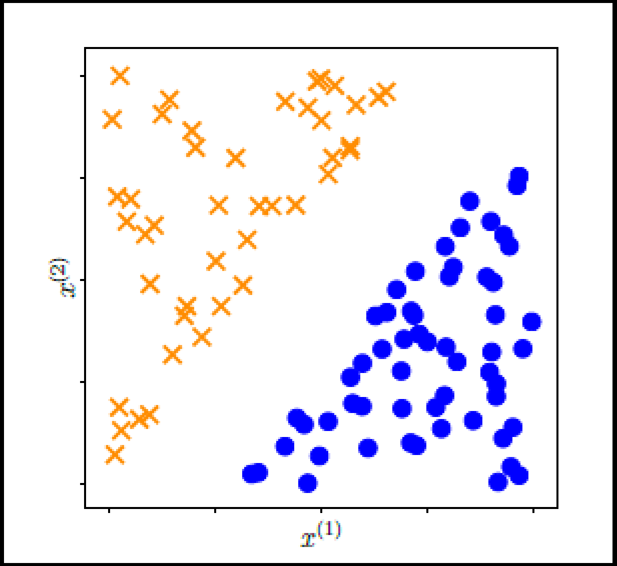
$$\lambda_1 = 0, \lambda_2 = 0, \lambda_3 = -2, \lambda_4 = -4, \lambda_5 = 0$$

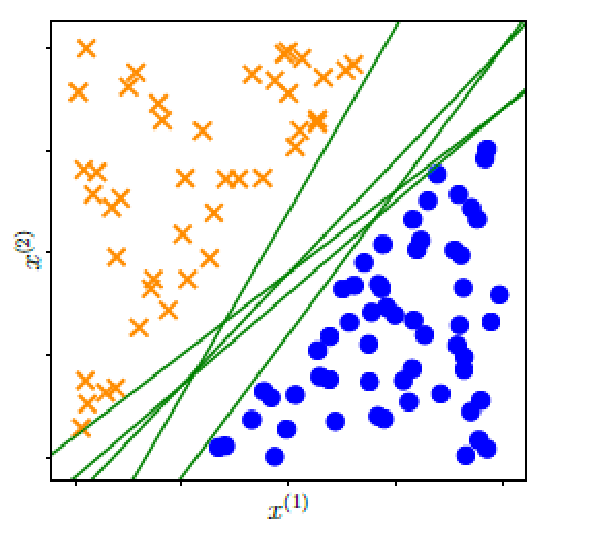
Since both $f(\boldsymbol{X})$ and the solution space $\boldsymbol{g}(\boldsymbol{X}) \leq \mathbf{0}$, the Lagrangian must be convex and the resulting stationary points yield a global constrained minima.

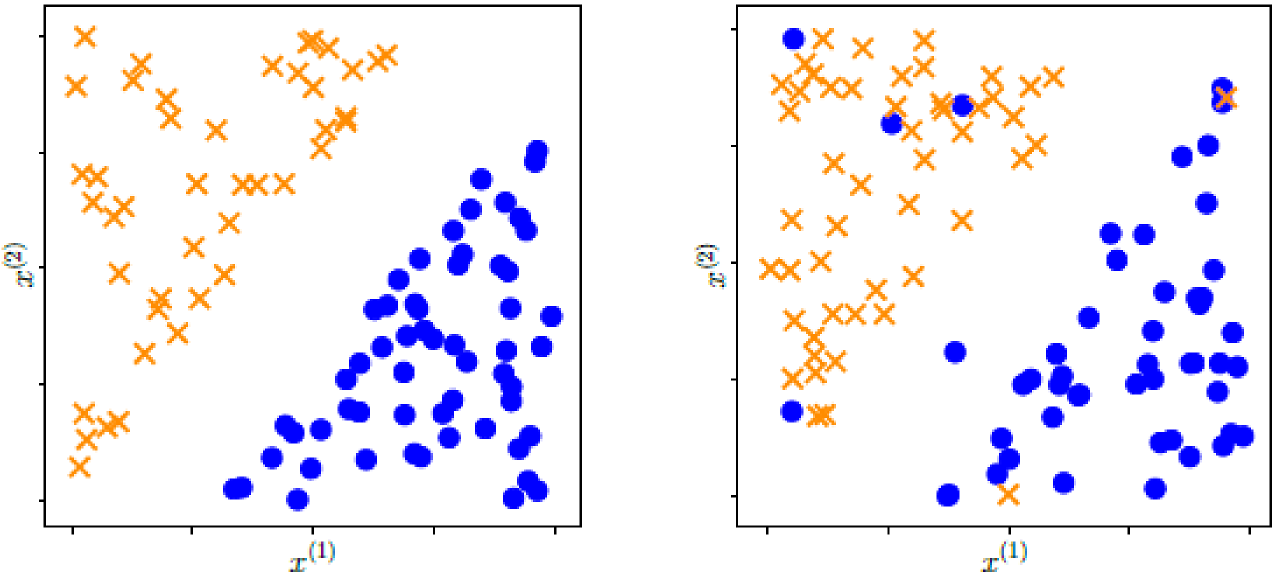
Applications to Machine Learning

Classification Problems: binary or n-ary

Support Vector Machine

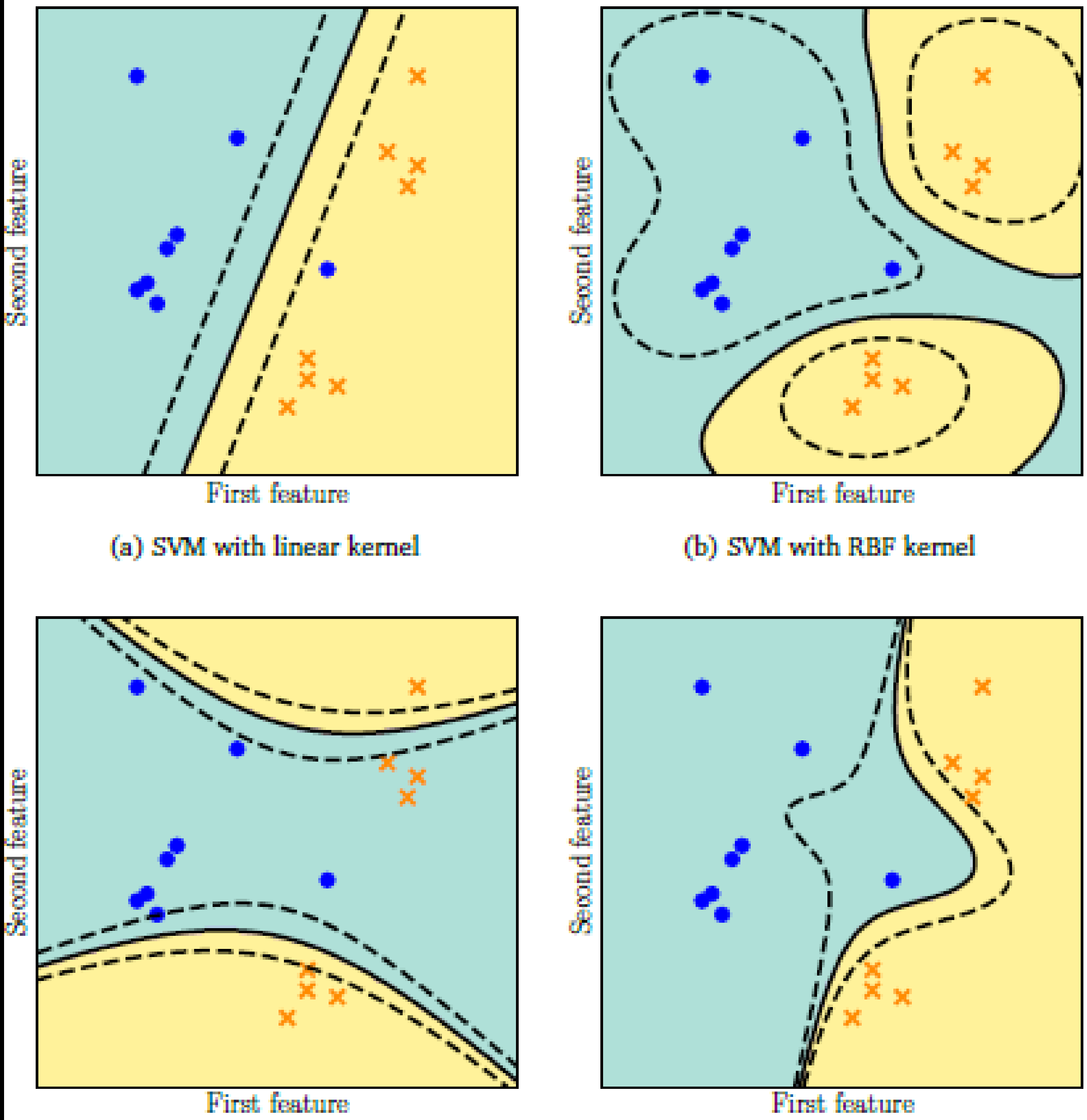






(a) Linearly separable data, with a large margin

(b) Non-linearly separable data



(a) *SVM* with linear kernel

(b) *SVM* with RBF kernel

(c) *SVM* with polynomial (degree 2) kernel

(d) *SVM* with polynomial (degree 3) kernel

1. Loss view of SVM can be expressed as an unconstrained optimization problem.

2. constrained versions of SVM can be expressed as quadratic programming problems

Numerical optimization

Gradient descent

Steepest descent

Optimization Using Gradient Descent

Let $f: \mathbb{R}^n \to \mathbb{R}$, such that f is differentiable, and it is not possible to find an analytical solution in closed form.

Consider: $\min_{\boldsymbol{X}} f(\boldsymbol{X})$

Gradient descent is a first-order optimization algorithm, which finds the local minima.

Gradient Descent

Def: A direction $\boldsymbol{d}$ is said to be a **direction of descent** of a function f at **X** if there exists a $\delta > 0$ such that $f(\boldsymbol{X} + \lambda\boldsymbol{d}) < f(\boldsymbol{X})$ for all $0 < \lambda < \delta$.

In particular, if in the limiting case:

$$\lim_{\lambda \to 0} \frac{f(\boldsymbol{X} + \lambda \boldsymbol{d}) - f(\boldsymbol{X})}{\lambda} < 0$$

then $\boldsymbol{d}$ is the **direction of descent.**

Observations:

- The gradient Descent Algorithm takes steps proportional to the negative of the gradient of the function at the current point.
- The gradient points in the direction of the steepest ascent.
- Consider the set of lines where the function is at a specified value $f(x) = c$, where c is a real number, called the contour lines. Then, the gradient points in a direction that is orthogonal to the contour lines.

Let initial guess be: X_0

Then, $f(X_0)$ will move <u>fastest</u>, if it moves in the direction of negative gradient, i.e. $-(\nabla f(X_0))^T$

If $X_1 = X_0 - \lambda (\nabla f(X_0))^T$

then $f(X_1) \leq f(X_0)$, provided $\lambda \geq 0$ is a small step size.

Gradient Descent Algorithm

To minimize $f(X)$, compute $\nabla f(X)$

Let X_0 be the initial guess

$$X_{k+1} = X_k - \lambda\,(\nabla f(X_k))^T$$

Where k is the iteration number.

This will give a sequence:

$f(X_0) \geq f(X_1) \geq f(X_2) \ldots \ldots ,$

which will converge to the local minima.

Example:

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 1 & 20 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 5 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Gradient=$\nabla f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 1 & 20 \end{bmatrix} - \begin{bmatrix} 5 & 3 \end{bmatrix}$

Take $X_0 = \begin{bmatrix} -3 \\ -1 \end{bmatrix}$. Then, $\nabla f(X_0) = \begin{bmatrix} -12 \\ -26 \end{bmatrix}$

Then $\quad X_1 = X_0 - \lambda \nabla f(X_0)$

$$= \begin{bmatrix} -3 \\ -1 \end{bmatrix} - 0.085\begin{bmatrix} -12 \\ -26 \end{bmatrix}$$

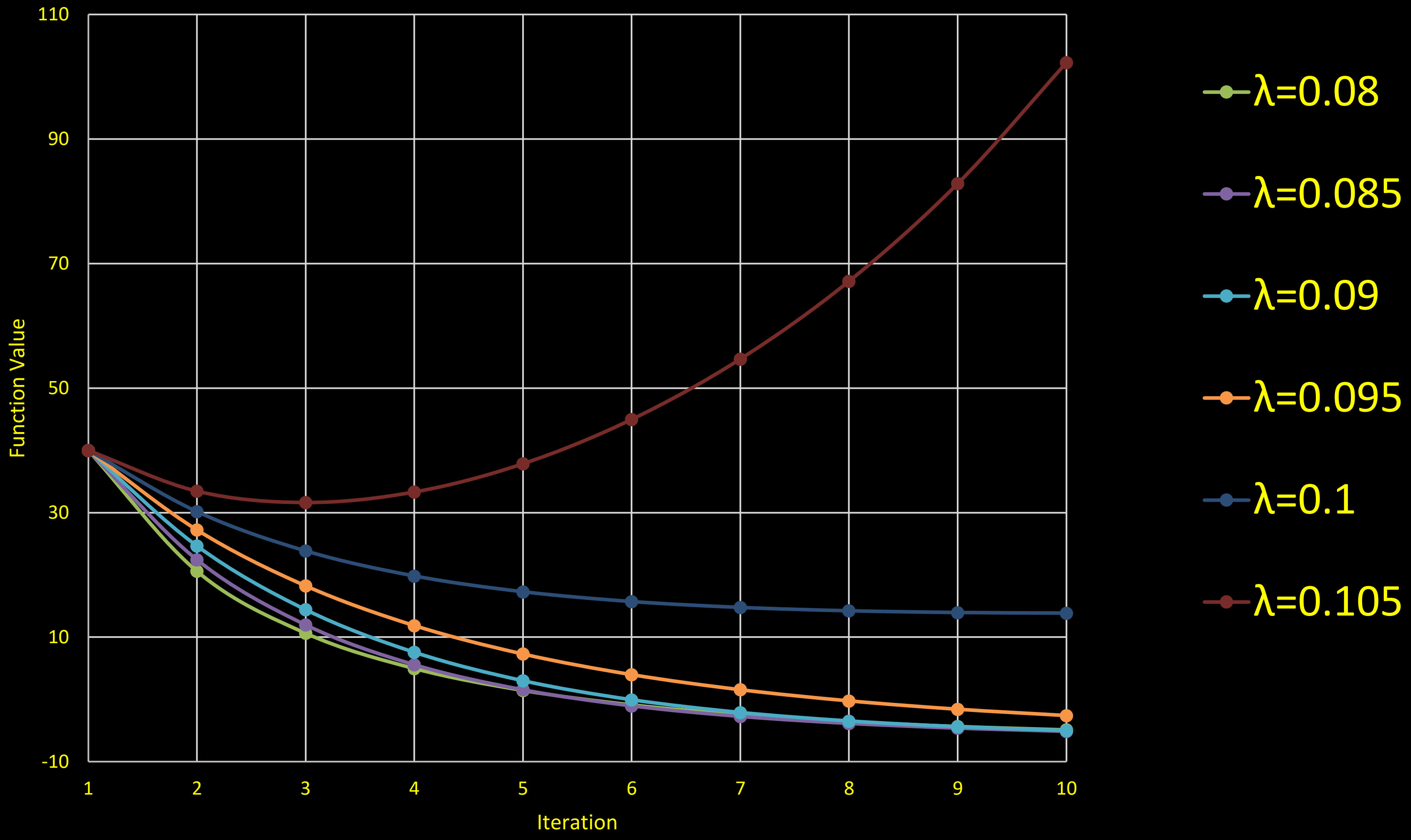
$$= \begin{bmatrix} -1.98 \\ 1.21 \end{bmatrix}$$

```
%% Gradient descent Method
X=[-3,-1];
A=[2,1;1,20];
B=[5,3];
max_iter=10;
lamda=0.085;

for i=1:max_iter
    f=1/2*X*A*transpose(X)-B*transpose(X);
    df=X*A-B;
    X=X-lamda*df;
    disp(['Iteration:',num2str(i)]);
    disp(['X is : ', num2str(X)]);
    disp(['Function value is : ', num2str(f)]);
    disp(['Gradient is : ', num2str(df)]);
end

%% Coutour graph
x=-5:1:5;
y=-5:1:5;
[X,Y]=meshgrid(x,y);
Z=0.5*(2*X.^2+23*X*Y+21*Y.^2+X+Y)-5*X-3*Y;
contour(X,Y,Z,'ShowText','on');
hold on;
meshgrid off
```

Function value for different λ

Iteration	λ=0.07	λ=0.075	λ=0.08	λ=0.085	λ=0.09	λ=0.095	λ=0.1	λ=0.105
0	40	40	40	40	40	40	40	40
1	17.9584	19.09	20.5824	22.4356	24.6496	27.2244	30.16	33.4564
2	10.1337	10.1004	10.6364	11.9781	14.4049	18.2392	23.846	31.6335
3	5.701	5.1344	4.9691	5.576	7.5638	11.8623	19.8208	33.3176
4	2.6225	1.9162	1.4484	1.5545	2.9928	7.2922	17.2815	37.8553
5	0.3625	-0.3289	-0.8718	-1.0293	-0.0629	3.9829	15.7073	44.9737
6	-1.318	-1.9408	-2.4569	-2.7213	-2.107	1.5606	14.7602	54.6723
7	-2.5712	-3.1101	-3.5617	-3.8464	-3.4749	-0.232	14.2207	67.1616
8	-3.5062	-3.9615	-4.3403	-4.6036	-4.391	-1.573	13.9466	82.8325
9	-4.204	-4.5822	-4.892	**-5.1179**	-5.0047	-2.587	13.8456	102.2479

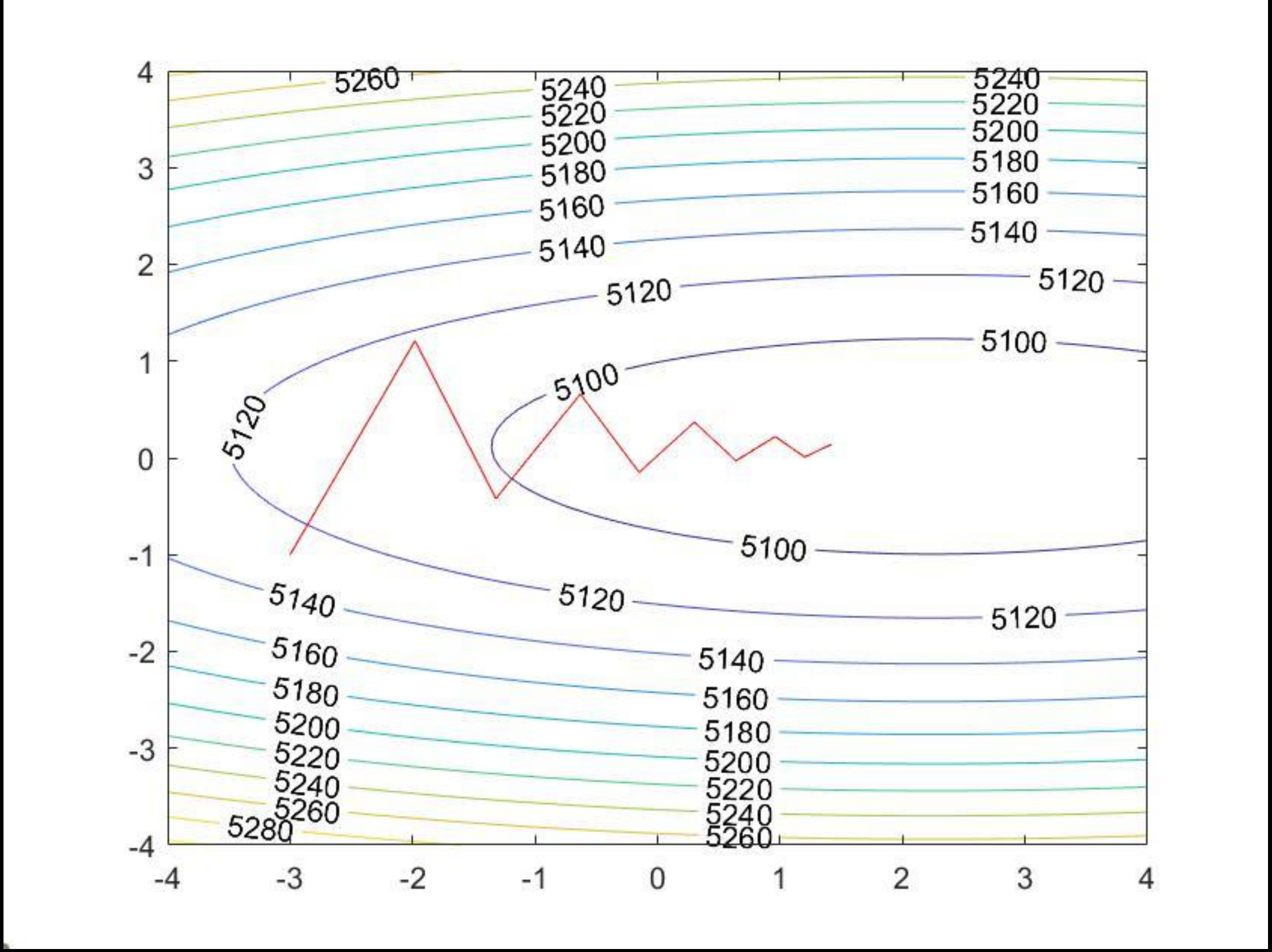
Iteration Vs Function Value for different λ



x_1 and x_2 for $\lambda = 0.085$

x_1	x_2	df/dx_1	df/dx_2	f
-3	-1	-12	-26	40
-1.98	1.21	-7.75	19.22	22.4356
-1.3213	-0.4237	-8.0662	-12.7953	11.9781
-0.6356	0.6639	-5.6073	9.6423	5.576
-0.159	-0.1557	-5.4737	-6.273	1.5545
0.3063	0.3775	-4.01	4.8564	-1.0293
0.6471	-0.0353	-3.7411	-3.0586	-2.7213
0.9651	0.2247	-2.8451	2.459	-3.8464
1.2069	0.0157	-2.5704	-1.4795	-4.6036
1.4254	0.1414	-2.0077	1.2541	-5.1179

Contour map showing convergence of $X_0, X_1, \ldots$



Strategies for Step size or learning rate

- Large in the earlier iterations and small in the later iterations
- Random adaptation
- Stochastic
- Steepest

Def: The direction of the negative gradient is the **steepest descent direction.**

Ex: $f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 1 & 20 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 5 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

Gradient=$\nabla f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 1 & 20 \end{bmatrix} - \begin{bmatrix} 5 & 3 \end{bmatrix}$

$$= \begin{bmatrix} 2x_1 + x_2 - 5 \\ x_1 + 20x_2 - 3 \end{bmatrix}$$

Then the steepest detection is:

$$-\nabla f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 5 - 2x_1 - x_2 \\ 3 - x_1 - 20x_2 \end{bmatrix}$$

Evaluated at $x_1 = x_2 = 0$, it is $\begin{bmatrix} 5 \\ 3 \end{bmatrix}$

Method of Steepest Descent

Initial Step:

Let $\in > 0$ be desired accuracy.

Choose a starting point X_0. Set $k = 0$.

Main Step

If $\|\nabla f(X_k)\| < \in$, stop, Otherwise let $d_k = -\nabla f(X_k)$.

Let λ_k be the optimal solution to the problem minimize $f(X_k + \lambda d_k)$ subject to $\lambda \geq 0$.

Set $X_{k+1} = X_k + \lambda_k d_{k\,.}$

Replace k by k + 1 and repeat main step.

Example:

$$f\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}\right)=\frac{1}{2}\begin{bmatrix}x_1 & x_2\end{bmatrix}\begin{bmatrix}2 & 1\\1 & 20\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}5 & 3\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}$$

$$\nabla f\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}\right)=\begin{bmatrix}x_1 & x_2\end{bmatrix}\begin{bmatrix}2 & 1\\1 & 20\end{bmatrix}-\begin{bmatrix}5 & 3\end{bmatrix}$$

Take $X_0=\begin{bmatrix}-3\\-1\end{bmatrix}, \nabla f(X_0)=\begin{bmatrix}-12\\-26\end{bmatrix}, d_0=\begin{bmatrix}12\\26\end{bmatrix}$

Let λ_0 = Minimize $f(X_0+\lambda d_0)$ subject to $\lambda\geq 0$

This is a one dimensional problem in λ.

$$X_0 + \lambda d_0 = \begin{bmatrix} -3 \\ -1 \end{bmatrix} + \lambda \begin{bmatrix} 12 \\ 26 \end{bmatrix} = \begin{bmatrix} 12\lambda - 3 \\ 26\lambda - 1 \end{bmatrix}$$

Putting $x_1 = 12\lambda - 3$ and $x_2 = 26\lambda - 1$,

we get $f(X_0 + \lambda d_0)$= $f\left(\begin{bmatrix} 12\lambda - 3 \\ 26\lambda - 1 \end{bmatrix}\right) =$

$$\frac{1}{2}\begin{bmatrix} 12\lambda - 3 & 26\lambda - 1 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 1 & 20 \end{bmatrix}\begin{bmatrix} 12\lambda - 3 \\ 26\lambda - 1 \end{bmatrix} - \begin{bmatrix} 5 & 3 \end{bmatrix}\begin{bmatrix} 12\lambda - 3 \\ 26\lambda - 1 \end{bmatrix}$$

$$= (12\lambda - 3)(12\lambda - 8) + (26\lambda - 1)(260\lambda - 29)$$

$$+(12\lambda - 3)(26\lambda - 1)$$

$$= 7216\lambda^2 - 1236\lambda + 56$$

For minima, $\frac{df}{d\lambda} = 0 \Longrightarrow 2 * 7216\lambda = 1236 \Longrightarrow \lambda = 0.0856$

So, $X_1 = X_0 + \lambda d_0 = \begin{bmatrix} -3 \\ -1 \end{bmatrix} + 0.0856 \begin{bmatrix} 12 \\ 26 \end{bmatrix} = \begin{bmatrix} -1.9728 \\ 1.2256 \end{bmatrix}$

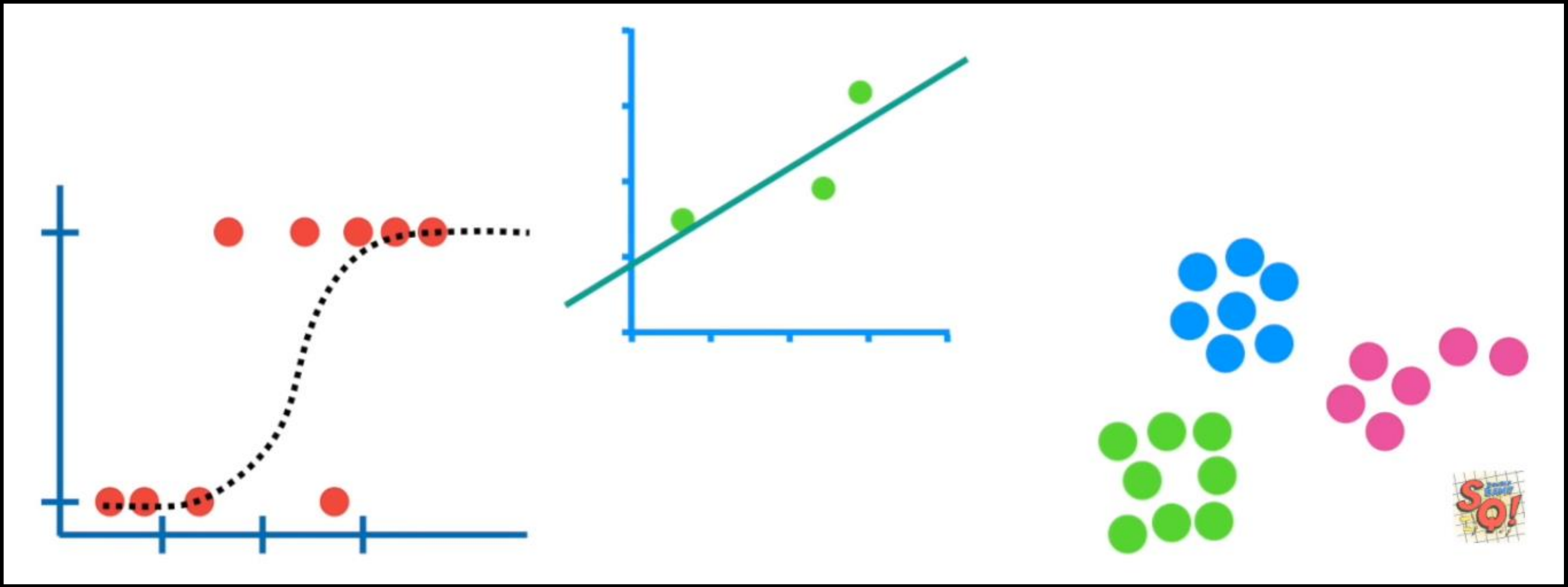
Iter k	$\boldsymbol{X_k}$ $\boldsymbol{f(X_k)}$	$\boldsymbol{\nabla f(X_k)}$	$\boldsymbol{\|\nabla f(X_k)\|}$	$\boldsymbol{d_k}$	$\boldsymbol{\lambda_k}$	$\boldsymbol{X_{k+1}}$
0	$\begin{bmatrix}-3\\-1\end{bmatrix}$	$\begin{bmatrix}-12\\-26\end{bmatrix}$	820	$\begin{bmatrix}12\\26\end{bmatrix}$	0.0856	$\begin{bmatrix}-1.9728\\1.2256\end{bmatrix}$
1	$\begin{bmatrix}-1.9728\\1.2256\end{bmatrix}$	$\begin{bmatrix}-7.72\\19.53\end{bmatrix}$	441.2	$\begin{bmatrix}7.72\\-19.53\end{bmatrix}$		
2						
3						

Other popular numerical techniques:

- Newton’s Method
- Conjugate Gradient Method
- Levenberg-Marquardt
- Golden section method
- Fibonacci Method
- ……………………

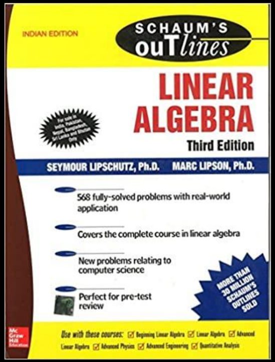
Applications in Machine Learning and Data Science

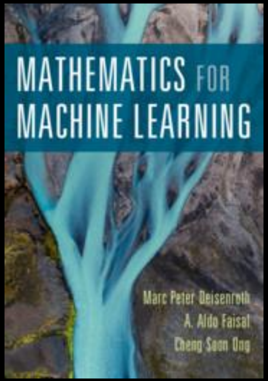
- Logistic Regression
- Linear Regression
- Clustering
- Many more….

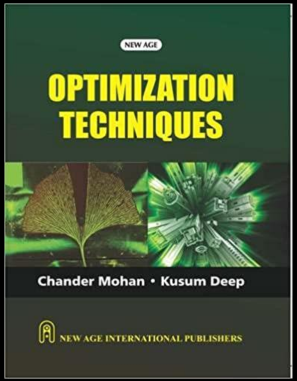


Recommended Books

1. Seymour Lipschutz and Marc Lars Lipson: Linear Algebra, Schaum's Outline, Third Edition, McGraw Hill, 2017.
2. Marc Peter Deisenroth, A. A. faisal, C. S. Ong: Mathematics for Machine Learning, Cambridge University Press, 2021.
3. Chander Mohan and Kusum Deep: Optimization Techniques, New Age Publications, 2nd Edition, 2023.







INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

श्रमं विना न किमपि साध्यम्

THANK YOU